Journal of Applied Mathematics and Physics, 2026, 14(1), 400-445
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Modeling and Forecasting Rental Prices in Berlin Using AI with Nonlinear Covariates for Housing Policy and Social Equity

Ugochukwu Onumadu[1*], Merci Iyelobu[2], Babatounde Yessoufou[3], Adedeji Adepeju[4], Sulaimon Adebayo[5], Oluwabusayo Omotosho[6]

[1]1Department of Educational Specialties, Austin Peay State University, Clarksville, USA
[2]Department of Economics, Northeastern University, Boston, USA
[3]The Fuqua School of Business, Duke University, Durham, North Carolina, USA
[4]Department of Architecture, Moshood Abiola Polytechnic, Abeokuta, Nigeria
[5]University of Massachusetts Amherst, Amherst, USA
[6]Department of Information Systems and Sciences, Bowie State University, Maryland, USA
Email: *uonumadu@my.apsu.edu, iyelobu.m@northeastern.edu, Babatounde.yessoufou@duke.edu, adepeju.adedeji@mapoly.edu.ng, sadebayo@umass.edu, oomotosho@bowiestate.edu

How to cite this paper: Onumadu, U., Iyelobu, M., Yessoufou, B., Adepeju, A., Adebayo, S. and Omotosho, O. (2026) Modeling and Forecasting Rental Prices in Berlin Using AI with Nonlinear Covariates for Housing Policy and Social Equity. *Journal of Applied Mathematics and Physics*, **14**, 400-445.
https://doi.org/10.4236/jamp.2026.141022

Received: November 21, 2025
Accepted: January 26, 2026
Published: January 29, 2026

Abstract

This study employs machine learning techniques (AI), specifically multiple linear regression with nonlinear covariates, to model rental prices per square meter in Berlin, Germany. The research investigates major quantitative and qualitative variables influencing rent dynamics by leveraging a robust dataset comprising over 2.6 million apartments with 59 variables from 2007 to 2020, sourced from FDZ Ruhr and ImmobilienScout24. Drawing from over 99,000 rental records (2015 and 2019 datasets) and 31 variables (9 quantitative and 22 qualitative), the analysis evaluates the influence of factors such as furnishing quality, modernization year, apartment size, and energy efficiency on rent pricing. Polynomial and log transformations were applied to improve model robustness. The use of log-transformed rent as the response variable, combined with nonlinear covariates, yielded the best model performance, with the highest adjusted Rsquared values of 0.3645 and 0.492 in the 2015 and 2019 models, respectively, among the tested models. The results suggest that both quantitative and qualitative variables significantly influence rent sqm, with influential predictors varying in linearity and significance across the two years. In 2015, rent was influenced by nonlinear trends in living space and construction years, while in 2019, heat cost and modernization showed linear increases in rent. Apartments with upscale furnishings, high energy efficiency, elevators, and guest toilets consistently commanded higher rents, whereas a pet allowance was associated with lower rents. Residual analysis, variance inflation factors (VIF), and the Akaike Information

Criterion (AIC) confirmed the model's statistical validity. Results indicate significant shifts in rent patterns and offer predictive insights relevant to policymakers, urban planners, and educational leaders addressing housing affordability and equity. This reseafrch builds prior literature in AI-supported urban analytics and contributes a replicable data-driven framework for strategic housing decisions in high-demand university cities.

Keywords

AI Rent Forecasting, Berlin Housing Market, Rental Price Modeling, Multiple Linear Regression, Social Equity, Policy Innovation

1. Background and Study Framework

1.1. Introduction

In this paper, artificial intelligence (AI) refers to the application of predictive modeling techniques, particularly regression analysis, within decision systems that use data and digital tools to support urban planning. Advanced statistical methods are used to model the rent per square meter (**rent_sqm**) in Berlin's housing market. Berlin's 2024 housing market exhibits a significant imbalance, with demand outpacing supply. In 2023, the population increased by 0.7% to 3.878 million, driven by 187,971 new arrivals, making it the third-highest year of immigration since 1991 [1]. Rental demand is high in Germany, particularly in Berlin and Munich, where prices have risen significantly, presenting challenges like those faced in other high-income countries, such as the UK, France, the US, and Canada [2]. Germany has a significantly higher share of renters compared to its international counterparts. In 2018, Germany's homeownership rate stood at 51.5%, while the rates in the UK, Italy, and Romania were 65.1%, 72.4%, and 96.4%, respectively [2]. Although AI often involves complex algorithms, it also encompasses predictive modeling for pattern recognition and analysis. This study uses a transformed multiple linear regression with nonlinear variables to forecast rent prices in Berlin, contributing to urban analysis efforts that support thoughtful policy decisions and promote fairness in housing [3].

1.2. Objective

This paper employs multiple linear regression to model rent_sqm in Berlin, examining major market trends, assessing the need for transformation of the response variable, and identifying significant predictors. The results aim to support the development of housing policy and promote equity, while also assisting educational leaders in strategic and equitable housing planning.

1.3. Literature Review

Skewness and variability in housing data are often addressed using log-linear regression, which highlights furnishing quality, energy efficiency, and modernization

as main influencing factors [4] [5]. Germany's public housing contrasts with the U.S.'s system based on market behavior [6]. Energy-efficient design has garnered increasing attention to sustainability [7]. This study employs AI-driven modeling to analyze Berlin's rental market and connect statistical insights to educational policy, with a focus on policy innovation and social equity in housing [8] [9].

1.4. Research Questions

- **RQ1:** Does a relationship exist between the response variable (rent_sqm) and the selected predictor variables?
- **RQ2:** Is a transformation of response variable (rent_sqm) necessary to meet the assumptions of linear regression?
- **RQ3:** Which predictor variables significantly affect the rental price per square meter in Berlin's housing market?

2. Methodology

This section explores the methodology and required mathematical and statistical background for this paper. The study uses multiple regression analysis to model rent trends in Berlin. Although regression is a classical statistical method, its application in this context serves as a robust AI tool (machine learning algorithms) for predictive modeling in housing policy. We look at linear models, their formulations, assumptions, estimations, validation, predictions, and hypothesis testing. More details can be found in [10]. The data analysis was executed using the R programming language.

2.1. Research Design

This research uses a quantitative approach and applies AI-informed predictive modeling, focusing on multiple linear regression with nonlinear covariates, to analyze Berlin's rental prices.

2.2. Data Collection

A secondary source of data collection was used for this study. The data was provided by the FDZ Ruhr at RWI (and ImmobilienScout24) institution. The Immobilien-Scout24 GmbH, founded in 1998, deals with real estate properties in Germany. The data set contains 2,651,885 observations and 59 attributes from 2007 to 2020. We selected Berlin City as it had the highest number of rental transactions in Germany. Thereafter, we chose the years 2015 and 2019, which had 49,724 and 49,536 records, respectively, based on the significant impact observed in the plotted scatter of years with rent prices as shown in the data description in **Table 3**.

2.3. Data Analysis: Multiple Linear Regression with Nonlinear Covariates

We cleaned the data and removed outliers using the interquartile range (IQR) method. The missing values recorded and the NAs were part of the labels for most categorical variables, as shown in **Table 2**. Exploratory data analysis (EDA) was

used to visualize the behavior of rent per square meter and its covariates. The distribution of rent prices was first explored through histograms, overlaid with fitted Normal and Log-Normal density functions to assess its shape as shown in Figure 1 and Figure 2. Scatter and box plots were used to observe the influential variables of rent price and how each variable enters the model. Eight models were fitted using multiple linear regression, and residual analysis, variance inflation factors (VIF), the Akaike information criterion (AIC), and the adjusted R-squared were used for model evaluation. Among the eight models, multiple linear regression with nonlinear covariates was used for the prediction of rent per square meter based on model evaluation.

2.3.1. Concept of a Multiple Linear Regression Model

Researchers frequently explore whether a relationship exists between variables and how they are connected if one is found. Regression models the relationship between a response variable and one or more independent variables (covariates); see [11]. For example, we might examine the relationship between apartment size (independent variable) and monthly rent (dependent or response variable). Regression analysis aims to estimate the parameters of the linear function that best describes the joint distribution of the dependent variable and the covariates [12]. Relationships between variables can be linear, nonlinear (e.g., quadratic or cubic), or absent. Exploratory Data Analysis (EDA) helps identify suitable model structures before fitting regression models. When multiple independent variables are used to predict a continuous response, the approach is called multiple linear regression. In this study, we aim to investigate the relationship between the rent per square meter in Berlin charged for an apartment, characterized by both continuous and discrete covariates.

2.3.2. Concept of a Multiple Linear Regression Model

Researchers frequently explore whether a relationship exists between variables and how they are connected if one is found. Regression models the relationship between a response variable and one or more independent variables (covariates); see [11]. For example, we might examine the relationship between apartment size (independent variable) and monthly rent (dependent or response variable). Regression analysis aims to estimate the parameters of the linear function that best describes the joint distribution of the dependent variable and the covariates [12]. Relationships between variables can be linear, nonlinear (e.g., quadratic or cubic), or absent. Exploratory Data Analysis (EDA) helps identify suitable model structures before fitting regression models. When multiple independent variables are used to predict a continuous response, the approach is called multiple linear regression. In this study, we aim to investigate the relationship between the rent per square meter in Berlin charged for an apartment, characterized by both continuous and discrete covariates.

2.3.3. Model Formulation

In a regression analysis with a continuous response variable Y_i and p covariates or predictors $X_{i1}, X_{i2}, \cdots, X_{ik}$ which may be continuous or qualitative (ordinal or

nominal) with n observations, let $\left(y_i, \boldsymbol{x}_i^\top\right) := \left(y_i, x_{i1}, \cdots, x_{ik}\right)^\top$, $i = 1, \cdots, n$, $k = p - 1$, be a pair of the ith observation $\left(y_i, \boldsymbol{x}_i^\top\right)$ of the random vector $\left(Y_i, \boldsymbol{x}_i^\top\right)$, where $\boldsymbol{x}_i = \left(x_{i1}, x_{i2}, \cdots, x_{ik}\right)^\top$, then our objective is to analyze the effects of the covariates on the mean value of the response variable ($\mu_i \equiv E\left[Y_i\right]$). The linear model models the response as a linear function of the predictors together plus an error term, i.e.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \epsilon_i \tag{2.1}$$

with mean $E\left[Y_i\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$.

Multiple linear regression model: The multiple linear regression model is defined as

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \ i = 1, \cdots, n, \tag{2.2}$$

where ε_i is the random error variable, β_0 is the intercept, and the k parameters $\beta_1, \cdots, \beta_k$ are the unknown regression parameters to be estimated from n observations $\left(y_i, x_{i1}, \cdots, x_{ik}\right)$, for $i = 1, \cdots, n$.

2.3.4. Matrix Representation

It is very easy to represent our linear model in a matrix form [13]. The four different model components below, are defined in order to represent the multiple linear regression model of (2.2) in the matrix-vector notation for our model formulation and calculation.

a) Let $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^n$, be the vector of the response variables.

b) Let $X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \in \mathbb{R}^{n \times p}$, be the design matrix that contains $p = k + 1$ predictors with their n observations in its rows. The columns correspond to the p unknown regression parameters. Note that the first column which corresponds to the intercept β_0 equals 1 for all n entries.

Denote by $\boldsymbol{x}_i \in \mathbb{R}^p$ the ith row of the design matrix.

c) Let $\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^p$, be the unknown vector of the regression coefficients.

d) Let $\boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \in \mathbb{R}^n$, be the vector of random error variables.

(*Linear regression model in matrix-vector notation*) With these notations

model (2.2) can be expressed as

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{n\times 1} = \underbrace{\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}}_{n\times p} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{p\times 1} + \underbrace{\begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}}_{n\times 1},$$

Thus, the multiple linear regression (2.2) can now be written as

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \epsilon_i, \text{ with } \epsilon_i \sim N_n\left(0, \sigma^2 I_n\right) \tag{2.3}$$

where I_n is the $n\times n$ identity matrix and $N_m\left(\boldsymbol{\mu}, \Sigma\right)$ denotes the m-dimensional multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

2.3.5. Assumptions of the Linear Model

a) **Linearity in the covariates**: In (2.1), we introduced that the relationship between the covariate vector $\boldsymbol{x}_i$ and the random response Y_i has the form

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i$$

with random error variable ε_i satisfying $E\left[\varepsilon_i\right] = 0$, so that

$$E\left[Y_i\right] = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

$$i = 1, \cdots, n.$$

In matrix notation: $E\left[\boldsymbol{Y}\right] = \boldsymbol{X\beta}$.

b) **Homoscedasticity**: The error variables ε_i have constant variance

$$\mathrm{Var}\left[Y_i\right] = \mathrm{Var}\left[\varepsilon_i\right] = \sigma^2, \ i = 1, \cdots, n.$$

c) **Independence of the random errors**: We assume that the error variables ε_i are independent and identically distributed (i.i.d.). Then it follows

$$\mathrm{Cov}\left(Y_j, Y_{j'}\right) = \mathrm{Cov}\left(\varepsilon_j, \varepsilon_{j'}\right) = 0, \forall j \neq j'$$

d) **Normality**: The random error variables ε_i follow a normal distribution.

$$\boldsymbol{Y} \sim N_n\left(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n\right) \tag{2.4}$$

where I_n is the $n\times n$ identity matrix and $N_m\left(\boldsymbol{\mu}, \Sigma\right)$ denotes the m-dimensional multivariate normal distribution with the mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, respectively.

In general, we assume a Gaussian error $\boldsymbol{\varepsilon}_i \sim N_n\left(0, \sigma^2 \boldsymbol{I}_n\right)$. This allows us to construct confidence intervals and conduct statistical tests.

Here, $\boldsymbol{I}_n$ is the n-dimensional identity matrix and $N_m\left(\boldsymbol{\mu}, \Sigma\right)$ denotes the m-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, respectively.

2.3.6. Polynomial Regression

Polynomial regression is often appropriate when there exists a relationship between the response and the covariates.

(*Polynomial regression*). Given a continuous covariate V_i with observations v_i that has a polynomial effect of degree d on the response, then the model $Y_i = \beta_0 + \beta_1 V_i + \beta_2 V_i^2 + \cdots + \beta_d V_i^d + \cdots + \varepsilon_i$ can be used. Note, it is a linear regression model of the form (2.2) with $x_{ij} = v_i^j$, $j = 1, \cdots, d$ [14] and [15].

In order to increase numerical stability, we orthonomalize the corresponding design matrix $X = \begin{pmatrix} 1 & v_1 & v_1^d \\ \vdots & \vdots & \vdots \\ 1 & v_n & v_n^d \end{pmatrix}$ to X^*, where all columns have unit norms and are orthogonal. In R, this is achieved by $\text{poly}(v, d)$, see [16].

2.3.7. Transformations of the Response Variable

Sometimes, the transformation of the response variable is appropriate when non-normality and/or unequal error variances are present in the data. We will consider three different transformations of the response variable in this paper.

(*Logarithmic and inverse transformation*) Given a response variable $\boldsymbol{Y}$ that has an exponential relationship with the covariates. Let $Y_i^{ln} := \ln(Y_i)$, let $Y_i^{lnln} := \ln(\ln(Y_i))$, let $Y_i^{inv} := \frac{1}{Y_i}$, then the formulated model $Y_i = \exp(\beta_0 + \beta_1 x_{i1}, \cdots, \beta_k x_{ik} + \varepsilon_i)$ can be expressed in the form of the linear regression model (2.2) as

$$Y_i^{ln} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \ i = 1, \cdots, n \tag{2.5}$$

$$Y_i^{lnln} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \ i = 1, \cdots, n \tag{2.6}$$

$$Y_i^{inv} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \varepsilon_i, \ i = 1, \cdots, n \tag{2.7}$$

2.3.8. Estimation of Model Parameters

In this section, we will examine the methods for estimating the unknown parameters in the linear regression model defined in Equation (2.2). Our goal is to determine estimates

$$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \cdots, \hat{\beta}_k\right)^\top \in \mathbb{R}^p \tag{2.8}$$

and the error variance σ based on n observations. Here $\boldsymbol{\beta}$ is the unknown regression parameter vector.

Note that parameter **estimators**, which are random quantities are different from their realizations called **estimates**, which are determined by the values of the observations. We will consider two approaches, least squares (LS) estimation, and maximum likelihood (ML) estimation. These two estimation methods yield the same estimator if the assumptions of independence, homoscedasticity, and normality of errors are satisfied.

2.3.9. Least Squares Estimation Method

Let the fitted values of the Model (2.2) be given as

$$\begin{aligned}\hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik},\ i = 1, \cdots, n \\ &= \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}\end{aligned} \tag{2.9}$$

Also, let the residual denoted by $\hat{\boldsymbol{\varepsilon}} = \left(\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n\right)' \in \mathbb{R}^n$, which is the difference between the observed response values y_i and the corresponding fitted values of (2.11), be given as

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}, \tag{2.10}$$

where $\hat{\boldsymbol{y}} = \left(\hat{y}_1, \cdots, \hat{y}_n\right)^\top \in \mathbb{R}^n$ in the vector notation. Then, least squares minimizes the residual sum of squares (the sum of the squared deviations) of Equation (2.12).

(*Sum of squared deviations*) Given the data $\left(y_i, x_i\right), i = 1, 2, \cdots, n$, the sum of the squared deviations which is used in obtaining the estimates $\hat{\boldsymbol{\beta}}$ of Equation (2.10) for the unknown regression parameters $\boldsymbol{\beta}$ is given as

$$Q_{LS}\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{n}\left(y_i - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}} \tag{2.11}$$

In order to minimize $Q_{LS}\left(\boldsymbol{\beta}\right)$ (2.13), we take the partial derivative of $Q_{LS}\left(\boldsymbol{\beta}\right)$ with respect to $\boldsymbol{\beta}$ and set the result to zero. Then, it follows

$$\frac{\partial\left(Q_{LS}\left(\boldsymbol{\beta}\right)\right)}{\partial\boldsymbol{\beta}} = \boldsymbol{0} \Leftrightarrow -2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} + 2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{0} \Leftrightarrow X^\top X\boldsymbol{\beta} = X^\top \boldsymbol{y} \tag{2.12}$$

We are now interested in solving the least squares normal equations given in (2.14). If the matrix X has a full rank p, then $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ will be positive definite and will have a unique solution. Thus, the minimum of $Q_{LS}\left(\boldsymbol{\beta}\right)$ is attained at

$$\hat{\boldsymbol{\beta}}_{LS} = \left(X^\top X\right)^{-1} X^\top \boldsymbol{y} \tag{2.13}$$

which is the least squares estimate from the normal equations.

2.3.10. Maximum Likelihood Estimation Method

The method of maximum likelihood estimation is based on specifying the distribution we are sampling from and writing the joint density of our sample, unlike in the least squares method where we do not specify the distribution of the response variable Y_i. Considering the assumptions of our linear model, we assumed in Equation (2.4) that the random variables Y_i are normally distributed ($\boldsymbol{Y} \sim N_n\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n\right)$). Thus, it follows that the likelihood of the vector $\left(\boldsymbol{\beta}, \boldsymbol{\sigma}\right)$ given the data values $\boldsymbol{y}$ is

$$L\left(\boldsymbol{\beta}, \boldsymbol{\sigma} \mid \boldsymbol{y}\right) = \frac{1}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right) \tag{2.14}$$

Therefore, the corresponding log likelihood is given by

$$l\left(\boldsymbol{\beta}, \boldsymbol{\sigma} \mid \boldsymbol{y}\right) = -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^{\mathrm{T}}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \tag{2.15}$$

To maximize this log-likelihood (2.17) with respect to $\boldsymbol{\beta}$, we differentiate Equation (2.17) with respect to $\boldsymbol{\beta}$ and set it equal to zero [17]. Thus, we have

$$\frac{\partial\left(l\left(\boldsymbol{\beta},\boldsymbol{\sigma}\,|\,\boldsymbol{y}\right)\right)}{\partial\boldsymbol{\beta}}=\mathbf{0}\ \Leftrightarrow\ -\frac{1}{2\sigma^2}\left(-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}+2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta}\right)=\mathbf{0}\ \Leftrightarrow\ X^{\top}X\boldsymbol{\beta}=X^{\top}\boldsymbol{y}\quad(2.16)$$

This shows that $\hat{\boldsymbol{\beta}}_{ML}=\hat{\boldsymbol{\beta}}_{LS}$.

Also, differentiating Equation (2.17) with respect to σ^2 and maximizing over σ^2, we have

$$\hat{\sigma}^2:=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\frac{1}{n}\left\|\hat{\boldsymbol{\varepsilon}}\right\|^2\quad(2.17)$$

2.3.11. Distribution of the Estimators

($\hat{\boldsymbol{Y}}$ and $\boldsymbol{H}$) We define the vector of the fitted random values $\hat{\boldsymbol{Y}}$ as

$$\hat{\boldsymbol{Y}}=X\hat{\boldsymbol{\beta}}=X\left(X^{\top}X\right)^{-1}X^{\top}\boldsymbol{Y}\quad(2.18)$$

Also, we define the hat matrix H which gives the projection of the vector $\boldsymbol{Y}$ onto the space that is spanned by the columns of the design matrix $\boldsymbol{X}$ as

$$H:=X\left(X^{\top}X\right)^{-1}X^{\top}\in\mathbb{R}^{n\times n}\quad(2.19)$$

It can be easily shown in Lemma (2.7.3) that H is both symmetric ($H^{\mathrm{T}}=H$) and idempotent ($H^2=H$) using the fact that $\left(X'X\right)^{-1}$ is symmetric ($\left[\left(X'X\right)^{-1}\right]'=\left(X'X\right)^{-1}$).

Lemma 2.1 (Symmetry and **Idempotence** of ***H*)** *Let* $H=X\left(X'X\right)^{-1}X'$. *Then* H *is symmetric and idempotent, i.e.,* $H'=H$ *and* $H^2=H$.

Proof.

$H'=\left(X\left(X'X\right)^{-1}X'\right)'=X\left[\left(X'X\right)^{-1}\right]'X'=X\left(X'X\right)^{-1}X'=H$, using $\left(AB\right)'=B'A'$.

$$HH=X\left(X'X\right)^{-1}X'X\left(X'X\right)^{-1}X'=H.$$

Since the estimators $\hat{\boldsymbol{\beta}}$ of the regression coefficients $\boldsymbol{\beta}$, the fitted values $\hat{\boldsymbol{Y}}$ and the raw residuals $\hat{\boldsymbol{\varepsilon}}$ are all linear functions of the vector of random variables Y_i, we can apply the transformation rules for expectation and variance-covariance matrix respectively, to show that

$$\begin{aligned}&E\left[\hat{\boldsymbol{\beta}}\right]=\boldsymbol{\beta},\quad &&\mathrm{Var}\left[\hat{\boldsymbol{\beta}}\right]=\sigma^2\left(X^{\top}X\right)^{-1},\\&E\left[\hat{\boldsymbol{Y}}\right]=\boldsymbol{X}\boldsymbol{\beta},\quad &&\mathrm{Var}\left[\hat{\boldsymbol{Y}}\right]=\sigma^2H,\\&E\left[\hat{\boldsymbol{\varepsilon}}\right]=0,\quad &&\mathrm{Var}\left[\hat{\boldsymbol{\varepsilon}}\right]=\sigma^2\left(I_n-H\right)\end{aligned}\quad(2.20)$$

Considering the normality assumption since $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{Y}}$, and $\hat{\boldsymbol{\varepsilon}}$ are linear functions of $\boldsymbol{Y}$, we have

$$\begin{aligned}&\hat{\boldsymbol{\beta}}\sim N_p\left(\boldsymbol{\beta},\sigma^2\left(X^{\top}X\right)^{-1}\right)\\&\hat{\boldsymbol{Y}}\sim N_n\left(\boldsymbol{X}\boldsymbol{\beta},\sigma^2H\right)\\&\hat{\boldsymbol{\varepsilon}}\sim N_n\left(0,\sigma^2\left(I_n-H\right)\right)\end{aligned}\quad(2.21)$$

It can be shown that the variance estimator $\hat{\sigma}^2$ given in (2.19) is given by

$$\mathrm{E}\left(\hat{\sigma}^2\right)=\frac{n-p}{n}\sigma^2.$$

and an unbiased estimator s^2 of σ^2 is given by

$$s^2 \coloneqq \frac{1}{n-p}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\frac{n}{n-p}\hat{\sigma}^2=\frac{1}{n-p}\left\|\hat{\boldsymbol{\varepsilon}}\right\|^2. \tag{2.22}$$

2.4. Goodness of Fit and Model Selection

It is of great importance to know the goodness of the fitted model after estimating the parameters of the linear regression model of (2.2). Thus, we need suitable measures of the goodness of fit. Therefore, we will introduce one of the appropriate measures of the goodness of fit called the coefficient of determination (R^2), which determines the proportion of variation of the response variable that is explained by the covariates.

2.4.1. Sum of Squares

(Sum of squares) We define the sum of squares SST **(total sum of squares)**, SSR **(regression sum of squares)** and SSE **(error sum of squares)** to quantify the amount of variability explained by the regression model as follows

$$\begin{aligned}
\mathrm{SST} &\coloneqq \sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2 \;\Leftrightarrow\; \left(\text{total sum of squares}\right)\\
\mathrm{SSR} &\coloneqq \sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2 \;\Leftrightarrow\; \left(\text{regression sum ofsquares}\right)\\
\mathrm{SSE} &\coloneqq \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \;\Leftrightarrow\; \left(\text{error sum of squares}\right)
\end{aligned} \tag{2.23}$$

where $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$. Thus, we can have the decomposition as

$$\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2+\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 \tag{2.24}$$

and using the fact that $\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)\left(y_i-\hat{y}_i\right)=0$, it follows from (2.26) that

$$\mathrm{SST}=\mathrm{SSR}+\mathrm{SSE} \tag{2.25}$$

2.4.2. Selection of Model (R^2 and Adjusted R^2)

The multiple coefficient of determination R^2 is a measure of goodness of fit. It measures how well the covariates in the model explain the variance in the response variable, see [18].

(Multiple coefficient of determination) We define the **multiple coefficient of determination** R^2 as

$$\mathrm{R}^2 \coloneqq \frac{\mathrm{SSR}}{\mathrm{SST}}=1-\frac{\mathrm{SSE}}{\mathrm{SST}} \tag{2.26}$$

We also define the **adjusted multiple coefficient of determination** $\mathrm{R}^2_{\mathrm{adj}}$ as

$$\mathrm{R}^2_{\mathrm{adj}} \coloneqq 1-\frac{\mathrm{SSE}/(n-p)}{\mathrm{SST}/(n-1)} \tag{2.27}$$

The values of the multiple coefficient of determination range from zero to one ($0 \le R^2 \le 1$). Our model accounts for a larger amount of variation of the response when the R^2 is closer to 1. However, the weakness of R^2 is that, it always increases when we add more covariates to our model, and therefore cannot be used to compare the goodness of fit for models with different numbers of covariates, see [19]. Thus, the need to establish an appropriate measure R^2_{adj} which compares models with different numbers of covariates. We will therefore make use of the adjusted multiple coefficient of determination (R^2_{adj}) as a measure of our model selection in this paper.

2.4.3. Correlation Analysis

To measure the strength and direction of the linear relationship between two continuous variables, we use the correlation analysis. The most commonly used metric is the Pearson correlation coefficient, denoted by ρ for the population and r for the sample. It ranges from −1 to 1, where values close to 1 or −1 indicate strong positive or negative linear relationships, respectively, and values near 0 suggest no linear relationship.

The sample Pearson correlation coefficient between two variables X and Y is given by:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\sqrt{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}},$$

where $\bar{X}$ and $\bar{Y}$ are the sample means of X and Y, respectively. This metric provides a preliminary indication of potential multicollinearity when applied to predictor variables.

2.5. Hypothesis Testing

A statistical hypothesis is an assumption about the form of a population, which based on sample information from the population, seeks to support or reject this assumption. If there is evidence that the null hypothesis (hypothesis of no difference) denoted by H_0 is not true, then it is rejected and its alternative denoted by H_1 is accepted. Thus, a test of hypothesis is a rule or a procedure used for deciding whether to accept or reject H_0 or to determine whether the observed sample differs significantly from expected results under H_0 [20]. This concept can be extended in statistical inference for the model parameters of linear regression [21]. For instance, we may want to know if the response variable is significantly influenced by a particular set of covariate variables, which can be expressed in terms of linear combinations of the unknown regression parameters $\boldsymbol{\beta} = \left(\beta_0, \cdots, \beta_k\right)^{\top}$. We will use the Chi-square, F and the univariate t-distribution since the t-test and the F-test rely on quantities of these distributions.

(Chi-square distribution) A continuous random variable X is said to have a **Chi-square distribution** with parameter, ν, if its probability density function is given by

$$f_X(x|\nu)=\frac{2^{-\nu/2}}{\Gamma(\nu/2)}x^{\nu/2-1}\mathrm{e}^{-x/2},\nu>0,x>0$$

Here, ν is the degree of freedom, $\mathrm{E}(X)=\nu$, $\mathrm{Var}(X)=2\nu$. Thus, we say that X follows a Chi-square distribution with ν degree of freedom ($X\sim\chi^2_\nu$).

(F-distribution) A continuous random variable X is said to have an **F-distribution** with degrees of freedom (df) ν_1 and ν_2, if its pdf is given by

$$f(x)=\frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}x^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1 x}{\nu_2}\right)^{\frac{\nu_1+\nu_2}{2}}},\ x\geq 0. \tag{2.28}$$

If $X_1\sim\chi^2_{\nu_1}$ and $X_2\sim\chi^2_{\nu_2}$ and are independent, it follows in (2.30) that X is F-distributed with ν_1 and ν_1 df.

$$X=\frac{X_1/\nu_1}{X_2/\nu_2}\sim F_{\nu_1,\nu_2} \tag{2.29}$$

(Univariate t-distribution) A continuous random variable X is said to have a **Univariate t-distribution** with degree of freedom df ν, if its pdf is given by

$$f_\nu(x;\mu,\sigma^2):=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{(\pi\nu)}\sigma}\left\{1+\left(\frac{x-\mu}{\sigma}\right)^2\frac{1}{\nu}\right\}^{-\frac{\nu+1}{2}},\ \nu\geq 1 \tag{2.30}$$

$$\mathrm{E}(X)=\mu \text{ and } \mathrm{Var}(X)=\frac{\nu}{\nu-2}\sigma^2.$$

If $X_1\sim N(0,1)$ and $X_2\sim\chi^2_n$ and are independent, it can be shown in (2.32) that T has a t-distribution with ν df.

$$T=\frac{X_1}{\sqrt{\frac{X_2}{\nu}}}\sim t_\nu. \tag{2.31}$$

2.6. F-Test

(General testing problem) We define the general testing problem as

$$H_0: C\boldsymbol{\beta}=\boldsymbol{d} \text{ versus } H_1: C\boldsymbol{\beta}\neq\boldsymbol{d} \tag{2.32}$$

where matrix $C\in\mathbb{R}^{q\times p}$ with $\mathrm{rank}(C)=q$ and $\boldsymbol{d}\in\mathbb{R}^q$ is called the **general linear hypothesis**. Using the distribution of the estimated regression coefficients $\hat{\boldsymbol{\beta}}$ given in (2.23), if H_0 is true, it follows that

$$\hat{\boldsymbol{\vartheta}}=C\hat{\boldsymbol{\beta}}-\boldsymbol{d}\overset{H_0}{\sim}N_q\left(\mathbf{0},\sigma^2C\left(X^\top X\right)^{-1}C^\top\right). \tag{2.33}$$

We used the fact that

$$\begin{aligned}&\mathrm{E}\left(\hat{\vartheta}\right)=C\mathrm{E}\left(\hat{\boldsymbol{\beta}}\right)-\boldsymbol{d}=C\mathrm{E}(\boldsymbol{\beta})-\boldsymbol{d}=\mathbf{0}\\&\mathrm{Var}\left(\hat{\vartheta}\right)=C\mathrm{Var}\hat{\boldsymbol{\beta}}C^\top=\sigma^2C\left(X^\top X\right)^{-1}C^\top.\end{aligned} \tag{2.34}$$

Also, using the spectral decomposition for a specific covariance and considering the definition of χ^2-distribution, it can be shown that

$$\frac{1}{\sigma^2}\hat{\boldsymbol{\vartheta}}^{\top}\left(C\left(X^{\top}X\right)^{-1}C^{\top}\right)^{-1}\hat{\boldsymbol{\vartheta}} \overset{H_0}{\sim} \chi_q^2 \tag{2.35}$$

One can also show that $\frac{1}{\sigma^2}\hat{\boldsymbol{\vartheta}}^{\top}\left(C\left(X^{\top}X\right)^{-1}C^{\top}\right)^{-1}\hat{\boldsymbol{\vartheta}} \sim \chi_q^2$ and $\mathrm{SSE}/\sigma^2 \sim \chi_{n-p}^2$, and are independent χ^2 distributed. We therefore define the statistic under H_0 as

$$F = \frac{\hat{\boldsymbol{\vartheta}}^{\top}\left(C\left(X^{\top}X\right)^{-1}C^{\top}\right)^{-1}\hat{\boldsymbol{\vartheta}}\Big/q}{\mathrm{SSE}/(n-p)} \overset{H_0}{\sim} F_{q,n-p}. \tag{2.36}$$

If the null hypothesis $H_0 : C\boldsymbol{\beta} = \boldsymbol{d}$ holds (ie, $C\boldsymbol{\beta} - \boldsymbol{d} = 0$), then we will reject H_0 for large values of F since small value of $C\hat{\boldsymbol{\beta}} - \boldsymbol{d}$ is expected.

Let the least square estimate among those $\boldsymbol{\beta}$ vectors which satisfy $C\boldsymbol{\beta} = \boldsymbol{d}$ be denoted as $\boldsymbol{\beta}_{H_0}$, i.e. it minimizes (2.13) under the condition $\boldsymbol{C\beta} = \boldsymbol{d}$. We define the corresponding sum of squares error for the LS fit under H_0 as

$$\mathrm{SSE}_{H_0} := \left\| \boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{H_0} \right\|^2$$

It can also be shown that $\hat{\boldsymbol{\vartheta}}^{\top}\left(C\left(X^{\top}X\right)^{-1}C^{\top}\right)^{-1}\hat{\boldsymbol{\vartheta}} = \mathrm{SSE}_{H_0} - \mathrm{SSE}$ is true which allows us to give the general F-test.

(General F test in linear regression) We define the test statistic $\boldsymbol{F}$ under the null hypothesis $H_0 : C\boldsymbol{\beta} = \boldsymbol{d}$ versus $H_1 : C\boldsymbol{\beta} \neq \boldsymbol{d}$ in the linear regression model of (2.2) as

$$F = \frac{\left(\mathrm{SSE}_{H_0} - \mathrm{SSE}\right)\Big/q}{\mathrm{SSE}/(n-p)} \overset{H_0}{\sim} F_{q,n-p} \tag{2.37}$$

We reject H_0 against H_1 at level α if

$$F > F_{(1-\alpha),q,n-p} \tag{2.38}$$

Here, $F_{(1-\alpha),q,n-p}$ is the $(1-\alpha)$ quantile of an F distribution with q and $n-p$ df. The quantity $n-p$ is also called the **residual degree of freedom**. Thus, we can now summarize the F-test procedure for our model (2.2) as follows:

Hypothesis

$$H_0 : (\beta_1, \cdots, \beta_k) = \mathbf{0}$$
$$H_1 : (\beta_1, \cdots, \beta_k) \neq \mathbf{0}$$

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

(No significant relationship exists between rent sqm and the predictors.)

$$H_1 : \text{At least one } \beta_j \neq 0, \text{ for some } j \in \{1, 2, \cdots, k\}$$

(A significant relationship exists between rent sqm and at least one predictor.)

Test statistic = F, defined in (2.37)

Rejection Rule: Reject H_0 at level α, if $F > F_{(1-\alpha),q,n-p}$

2.7. T-Test

(t-test) We define the **t-test procedure** for our model (2.2) as follows, since in a t-test, the test statistic is computed for each β_j, see [22].

Hypotheses:

$H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$

Test statistic:

$$T_j = \frac{\hat{\beta}_j}{\widehat{se}\left(\hat{\beta}_j\right)} \sim t_{n-p}, \text{ under } H_0 \tag{2.39}$$

Here, $\widehat{se}\left(\hat{\beta}_j\right) := s\sqrt{\left(\left(X^\top X\right)^{-1}\right)_{jj}}$ is the estimated standard error of $\hat{\beta}_j$ and $s = \sqrt{s^2}$ defined in Equation (2.22)

Rejection Rule: Reject H_0 at level α, if $\left|T_j\right| > t_{n-p,1-\alpha/2}$

Note that **t-test** is a special case of the **F-test**, in particular we have $F_j = T_j^2$ since if

$$\begin{aligned} F_j &:= \frac{\hat{\beta}_j^2}{\left(\left(X^\top X\right)^{-1}\right)_{jj} \mathrm{SSE}/(n-p)} \overset{H_0}{\sim} F_{1,n-p} \\ T_j &:= \frac{\hat{\beta}_j}{\widehat{se}\left(\hat{\beta}_j\right)} \overset{H_0}{\sim} t_{n-p} \end{aligned} \tag{2.40}$$

2.8. Analysis of Variance (ANOVA)

ANOVA is mostly used to summarize the hypothesis tests results in linear models in a tabular form. Given two models M_{reduced} and M_{full} which are nested: $M_{\text{reduced}} \subset M_{\text{full}}$, that is, all covariates of the reduced model are contained in the full model, we define the **ANOVA-test ratio** for the comparison of M_{reduced} and M_{full} as follows

$$F = \frac{\left(\mathrm{SSE}_{\text{reduced}} - \mathrm{SSE}_{\text{full}}\right)/\left(n - p_{\text{full}}\right)}{\mathrm{SSE}_{\text{reduced}}/\left(p_{\text{full}} - p_{\text{reduced}}\right)} \sim F_{n-p_{\text{full}}, p_{\text{full}}-p_{\text{reduced}}} \tag{2.41}$$

Hypotheses

$$H_0 : \beta_i = 0 \text{ versus } H_1 : \beta_i \neq 0$$

Test statistic: F, defined in Equation (2.41)

Rejection Rule: Reject H_0 at level α, if $F > F_{(1-\alpha),n-p_{\text{full}},p_{\text{full}}-p_{\text{reduced}}}$

2.9. Analysis of Residuals

After estimating the model parameters, the credibility of the assumptions of linearity, normality of errors, and homoscedasticity for the given data can be assessed using residuals. It is therefore of importance to study the residual in order to examine in what extent our model assumptions may be violated. Therefore, taking a look at the patterns in the residual plots could help us understand if our

model assumptions are violated or not. This is called analysis of residual. Residual plots can equally help us to decide whether to transform any of the covariates which we may want to include in the model or not. We will introduce three types of residuals for the ith observation namely: **raw residuals**, **internally studentized residuals** and **externally studentized residuals**.

2.10. Raw Residuals Check

The **raw residual vector** $\hat{\boldsymbol{\varepsilon}}$ was defined in (2.12), and its distribution in (2.22). Thus, we have that

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{Y} - \underbrace{X\left(X'X\right)^{-1}X'}_{H}\boldsymbol{Y} = \boldsymbol{Y} - H\boldsymbol{Y} = \left(I_n - H\right)\boldsymbol{Y}$$

and

$$\mathrm{Var}\left(\hat{\boldsymbol{\varepsilon}}_i\right) = \sigma^2\left(1 - h_{ii}\right), \tag{2.42}$$

where $h_{ii} = \boldsymbol{x}_i^{\top}\left(X^{\top}X\right)^{-1}\boldsymbol{x}_i$ is the i th diagonal element of the hat-matrix H defined in (2.21). Since $\mathrm{Var}\left(\hat{\boldsymbol{\varepsilon}}_i\right)$ still changes under the homoscedasticity condition, we need to standardize the raw residuals. This standardization gives rise to both **internally studentized residuals** and **externally studentized residuals**.

Internally studentized residuals

The internally studentized residuals are defined by

$$s_i := \frac{\hat{\boldsymbol{\varepsilon}}_i}{\sqrt{1 - h_{ii}}s}, \tag{2.43}$$

Here, s^2 is the estimate of σ^2 defined in (2.19). One can now analyze the variances and conclude whether the assumption of homoscedasticity is violated or not using the standardized residuals by plotting the standardized residuals versus the predicted values $\hat{y}_i$. However, the deficiency of the internally studentized residuals is that it is not robust against outliers since all data is used to estimate σ and Equation (2.43) is t-distributed. This leads to the definition of externally studentized residuals which is more robust against outliers.

2.11. Externally Studentized Residuals or Jackknifed Residuals

To solve the problem of the non robustness of the internally studentized residuals, we define a new model just like the model of (2.3), but it is based on "drop-one-observation" of the data, which contains all observations except the i_{th} observation as

$$\boldsymbol{Y}_{-i} = X_{-i}\boldsymbol{\beta}_{-i} + \boldsymbol{\varepsilon}_{-i}, \tag{2.44}$$

Here, X_{-i} denotes the design matrix without the row ith and $\boldsymbol{Y}_{-i}$ is the response vector $\boldsymbol{Y}$ with the ith observation y_i removed. We define the fitted values corresponding to the model given in (2.44) as

$$\hat{y}_{i,-i} := \boldsymbol{x}_i^{\top}\hat{\boldsymbol{\beta}}_{-i} \tag{2.45}$$

where $\hat{\boldsymbol{\beta}}_{-i}$ is the associated least squares estimates of $\boldsymbol{\beta}_{-i}$. We define the

corresponding residual also called the ith **predictive residual** as

$$\hat{\boldsymbol{\varepsilon}}_{i,-1} := y_i - \hat{y}_{i,-i} \tag{2.46}$$

Using (2.46), we obtain the estimate s_{-i}^2 of the error variance σ^2 which does not include the ith observation as

$$s_{-i}^2 := \frac{\sum_{j=1,j\neq i}^{n}\left(y_j - \boldsymbol{x}_j^{\top}\hat{\boldsymbol{\beta}}_{-i}\right)^2}{n-p-1} \tag{2.47}$$

Finally, we now define the externally studentized residuals which is also called jackknifed residual. The externally studentized residuals t_i which is based on the "drop-one-observation" of the data, are defined as

$$t_i := \frac{r_{i,-i}}{\sqrt{1-h_{ii}}s_{-i}} \tag{2.48}$$

2.12. Statistical Checks for the Plausibility of the Linear Model Assumptions

2.12.1. Linearity

The check we are going to use is the residuals versus the fitted values plot. If this plot has no trend, then we assume the linearity assumption as plausible [23].

2.12.2. Homoscedasticity

We are interested in checking if $\mathrm{Var}\left[Y_i\right] = \mathrm{Var}\left[\varepsilon_i\right] = \sigma^2$ holds. To check this, we use again the standardized residual versus the residual plots. If the standardized residuals are not spread equally along the range of the fitted values, then we interpret the homoscedasticity assumption as not plausible, see [24].

2.12.3. Independence

To check if $\mathrm{Cov}\left(\varepsilon_j, \varepsilon_j'\right) = \rho = 0$ holds, we plot the residuals versus the covariates to see if the residuals are randomly and symmetrically distributed around zero. If this is true, we assume that the independence assumption is plausible [23].

2.12.4. Normality

To check for $\varepsilon_i \sim N_n\left(0, \sigma^2\boldsymbol{I}_n\right)$, we use the Quantile versus Quantile plot (QQPlot). If we do not have a straight line on the QQ plots of our variable versus the theoretical normal quantile, then we assume that the normality assumption is not plausible [25].

2.12.5. Multicollinearity

To check for multicollinearity among explanatory variables $X_1, X_2, \cdots, X_p$, we assess whether there is a strong linear relationship between them, which can inflate the standard errors of the estimated coefficients $\hat{\beta}_j$. This is commonly evaluated using the Variance Inflation Factor (VIF), defined as

$$\mathrm{VIF}_j = \frac{1}{1-R_j^2},$$

where R_j^2 is the coefficient of determination from regressing X_j on the

remaining predictors. A $\mathrm{VIF}_j > 5$ suggests a potentially problematic level of multicollinearity. Variables exceeding this threshold must be examined and removed if necessary to enhance model stability and interpretability. [26] [27].

3. Data Description and Management

The data set contains 2,651,885 observations and 59 attributes from 2007 to 2020. The data has both quantitative and qualitative covariates with **rent per square meter (rent_sqm)** as the response variable. We focus on the most relevant 31 variables such as "the additional cost", "heat cost", "construction year", etc. The quantitative covariates are summarized as follows: Min = Minimum, 25% = 1st quartile, 50% = Median, $\bar{X}$ = Mean, 75% = 3rd quartile, Max = Maximum and Not available = NA. On the other hand, the qualitative covariates are summarized with their respective categories. Note that costs are expressed in EUR and rounded to two decimal digits and the following data summaries in Table 1 and Table 2 represent the whole data set.

Table 1. Description of quantitative variables.

Variables	Description
rent_sqm	Calculated rent per sqm by rent and size of apartment. Min = 3, 25% = 7, 50% = 9, $\bar{X}$ = 9.39, 75% = 12, Max = 28
addcost	The extra monthly costs that need to be paid for other bills on top of the base rent excluding electricity. Min = 0, 25% = 100, 50% = 140, $\bar{X}$ = 153.8, 75% = 196, Max = 599, NA = 97186
heatcost	The monthly heating cost. Min = 0, 25% = 50, 50% = 70, $\bar{X}$ = 75.2, 75% = 94, Max = 300, NA = 898984
conyear	The year in which the object was built Min = 1851, 25% = 1930, 50% = 1970, $\bar{X}$ = 1964, 75% = 1996, Max = 2020, NA = 447372
lmod	The year of the last modernization Min=1981, 25% =2009, 50%=2012, $\bar{X}$ =2011, 75%=2015, Max=2018, NA=1113056
lspace	Living space in square meters Min = 19, 25% = 53, 50% = 68, $\bar{X}$ = 71.15, 75% = 85, Max = 165
fspace	The usable floor space in square meters Min = 0, 25% = 16, 50% = 57, $\bar{X}$ = 54.8, 75% = 79, Max = 250, NA = 1053922
energycon	The energy consumption per year and square meter in kWh Min = 0, 25% = 82, 50% = 117, $\bar{X}$ = 120.4, 75% = 152, Max = 350, NA = 977343
adlength	The difference between edat and adat. Min = 0, 25% = 0, 50% = 0, $\bar{X}$ = 0.71, 75% = 1, Max = 20

The descriptive statistics reveal notable variation across key quantitative variables relevant to rent modeling. Rent per square meter varies significantly, indicating differences in property value influenced by factors such as location, condition, and

amenities. Additional cost (adcost) and heating costs (heatcost) also show a wide spread, with some properties incurring substantially higher energy costs, possibly due to inefficiency. The construction (conyear) and modernization years (lmod) extend across several decades, which capture both historic and newly renovated buildings. Living (lspace) and floor space (fspace) differ widely, which shows a mix of compact and spacious apartments in the dataset. Energy consumption (energycon) values highlight a range of efficiency levels, which likely impact rental prices. The duration of the advertisement ranges from zero to several weeks, suggesting varying market demand and listing strategies. These differences confirm the complexity of accurately predicting rental prices. As a result, the study proposes the use of a log-normal model incorporating nonlinear covariates to better capture the heterogeneity and improve prediction accuracy in rent data.

Table 2. Description of qualitative variables.

Variables	Description
afloor	Apartment-specific variable indicates the floor the apartment is located in. **afloorg** is used to group afloor as follows: (−1) - 0, 1 - 2, 3 - 9, >9, NA
bfloor	This indicates the number of floors in the building. **bfloorg** is used to group bfloor as follows: 0 - 2, 3, 4, 5, >5, NA
nrooms	Number of rooms, excluding kitchen, bath or corridors. **nroomsg** is used to group nrooms as follows: 1 - 1.5, 2 - 2.5, 3 - 3.5, >3.5, NA
nbed	Number of bedrooms of the property. **nbedg** is used to group nbed as follows: 0 - 1, 2, >2, NA
nbath	Number of bathrooms of the property **nbathg** is used to group nbath as follows: 0 - 1, >1, NA
elevator	This variable indicates if a property has an elevator. **elevatorg** is used to group elevator as follows: Yes, No, NA
balcony	This variable indicates the presence of a balcony. **balconyg** is used to group balcony as follows: Yes, No, NA
kitchen	This variable indicates the presence of a fitted kitchen. **kitcheng** is used to group kitchen as follows: Yes, No, NA
eww	if the warm water consumption was included in the energy consumption value calculation. **ewwg** variable is used to group eww as follows: Yes, No, NA

Continued

subh	It indicates if a certificate of eligibility to public housing is needed to rent the apartment. **subhg** is used to group subh as follows: Yes, No, NA
gtoilet	This indicates the presence of a guest toilet. **gtoiletg** is used to group gtoilet as follows: Yes, No, NA
garden	This indicates the presence of a garden. **gardeng** is used to group garden as follows: Yes, No, NA
hww	if the warm water consumption was included in the heating cost value calculation. **hwwg** is used to group hww as follows: Yes, No, NA
cellar	This indicates if an property has a cellar room **cellarg** is used to group cellar as follows: Yes, No, NA
parking	This variable indicates whether a parking space is available. **parking** is used to group parking as follows: Yes, No, NA
furnishing	This is an artificial category number indicating the facilities of the property. **furnishingg** is used to group furnishing as follows: (Upscale, Luxury) = Upscale, (Normal, Simple) = Normal, no specification = NA
eeff	This indicates the energy efficiency rating. **eeffg** is used to group eeff as follows: (A, APLUS, B) = High, (C, D, E) = Medium, (F, G, H) = Low, no specification = NA
ecert	The type of energy performance certificate that the customer has for the object **ecertg** is used to group ecert as follows: Final energy demand = building, Energy consumption characteristic = consumption, NA
pets:	This indicates whether pets are allowed in the property. **petsg** is used to group pets as follows: (Yes, by Agreement) = Yes, No = No, no specification = NA
heat	This indicates the type of heating. **heatg** is used to group heat as follows: Central Heating (CH), Non Central Heating (NCH), NA
apcat	This variable categorizes the property into different classes. **apcatg** is used to group apcat as follows: (Penthouse, Maisonette, Attic Apartment) = top, Apartment = middle, (Mezzanine, Terrace apartment) = low, Basement = below, NA
pcon	This indicates the condition of a property. **pcong** is used to group pcon as follows: (First occupancy, First occupancy after renovation) = First, (Maintained, as good as new) = Mt, In need of renovation = Inr, (Modernized, Renovated, Fully Renovated) = Md, NA

3.1. Data Sets

We split the date set described in Table 1 and Table 2 into two sub data sets: Berlin 2015 and Berlin 2019. The number of rental properties contained in each data set is given in Table 3. The summaries of the response variable and the quantitative covariates are given in Table 4, while in Table 5, we give the summary of each qualitative variable followed by their percentage.

Table 3. Number of rental properties in the two data sets.

city	2015	2019
Berlin	49,724	49,536

Table 4 Univariate data summaries of quantitative covariates: first row = Berlin 2015, second row = Berlin 2019.

	Summary						
Variable	Min	25%	50%	Mean	75%	Max	NA
rent_sqm 2015	3.00	7.00	8.00	8.66	10.00	17.00	0
rent_sqm 2019	4.00	8.00	11.00	11.62	14.00	27.00	0
addcost							
	0.00	97.00	140.00	154.24	195.00	592.00	2021
	0.00	100.00	141.00	157.21	200.00	599.00	1070
heatcost							
	0.00	54.00	75.00	81.18	100.00	300.00	21126
	0.00	49.00	65.00	71.44	90.00	300.00	19077
conyear							
	1851	1910	1961	1954	1992	2016	7647
	1853	1918	1972	1964	1998	2020	6437
lmod							
	1983	2012	2014	2012	2015	2016	34384
	1982	2013	2016	2014	2018	2018	42152
lspace							
	23.00	55.00	69.00	73.80	89.00	161.00	0
	19.00	52.00	65.00	68.64	82.00	158.00	0
fspace							
	0.00	50.00	67.00	69.51	89.00	220.00	35475
	0.00	48.00	65.00	67.69	87.00	250.00	41068

Continued							
energycon							
	0.00	88.00	117.00	121.47	149.00	350.00	16665
	0.00	74.00	105.00	110.69	140.00	347.00	13863
adlength							
	0.00	0.00	0.00	0.71	1.00	20.00	0
	0.00	0.00	0.00	0.49	1.00	20.00	0

Table 5. Univariate data summaries of qualitative covariates: first row = Berlin 2015, second row = Berlin 2019.

Variable	Categories					
afloorg	(−1) - 0	1 - 2	3 - 9	>9	NA	
	4174 (0.08%)	18918 (0.38%)	19719 (0.4%)	868 (0.02%)	6045 (0.12%)	
	4320 (0.09%)	18759 (0.38%)	21016 (0.42%)	1020 (0.02%)	4421 (0.09%)	
bfloorg	0 - 2	3	4	5	>5	NA
	3222 (0.06%)	4938 (0.1%)	10313 (0.21%)	8974 (0.18%)	8030 (0.16%)	14247 (0.29%)
	2659 (0.05%)	4426 (0.09%)	8838 (0.18%)	9291 (0.19%)	8698 (0.18%)	15624 (0.32%)
nroomsg	1 - 1.5	2 - 2.5	3 - 3.5	>3.5		
	7283 (0.15%)	20391 (0.41%)	15612 (0.31%)	6438 (0.13%)		
	8959 (0.18%)	21660 (0.44%)	14465 (0.29%)	4452 (0.09%)		
nbedg	0 - 1	2	>2	NA		
	14796 (0.3%)	8465 (0.17%)	3710 (0.07%)	22753 (0.46%)		
	10951 (0.22%)	6107 (0.12%)	2204 (0.04%)	30274 (0.61%)		
nbathg:	0-1	>1	NA			
	28941 (0.58%)	3681 (0.07%)	17102 (0.34%)			
	25237 (0.51%)	3151 (0.06%)	21148 (0.43%)			
elevator:	Yes	No	NA			
	17145 (0.34%)	30648 (0.62%)	1931 (0.04%)			
	21019 (0.42%)	28517 (0.58%)	0 (0%)			
balconyg:	Yes	No	NA			
	33799 (0.68%)	15207 (0.31%)	718 (0.01%)			
	35112 (0.71%)	14424 (0.29%)	0 (0%)			
kitcheng:	Yes	No	NA			
	23510 (0.47%)	24111 (0.48%)	2103 (0.04%)			
	23390 (0.47%)	26146 (0.53%)	0 (0%)			

Continued

ewwg:	Yes	No	NA	
	12923 (0.26%)	36042 (0.72%)	759 (0.02%)	
	4701 (0.09%)	3449 (0.07%)	41386 (0.84%)	
subhg:	Yes	No	NA	
	1123 (0.02%)	43181 (0.87%)	5420 (0.11%)	
	3550 (0.07%)	45986 (0.93%)	0 (0%)	
gtoiletg:	Yes	No	NA	
	6404 (0.13%)	43237 (0.87%)	83 (0.00%)	
	4995 (0.10%)	44541 (0.90%)	0 (0%)	
gardeng	Yes	No	NA	
	6740 (0.14%)	39412 (0.79%)	3572 (0.07%)	
	5250 (0.11%)	44286 (0.89%)	0 (0%)	
hwwg:	Yes	No	NA	
	23355 (0.47%)	23785 (0.48%)	2584 (0.05%)	
	24338 (0.49%)	24114 (0.49%)	1084 (0.02%)	
cellarg:	Yes	No	NA	
	27227 (0.55%)	22194 (0.45%)	303 (0.01%)	
	24999 (0.50%)	24537 (0.50%)	0 (0%)	
parkingg:	Yes	No	NA	
	113 (0.00%)	2 (0.00%)	49609 (1.00%)	
	8133 (0.16%)	486 (0.01%)	40917 (0.83%)	
furnishingg:	Upscale	Normal	NA	
	15678 (0.32%)	11993 (0.24%)	22053 (0.44%)	
	14417 (0.29%)	8664 (0.17%)	26455 (0.53%)	
eeffg:	High	Medium	Low	NA
	796 (0.02%)	1084 (0.02%)	256 (0.01%)	47588 (0.96%)
	1434 (0.03%)	2039 (0.04%)	348 (0.01%)	45715 (0.92%)
ecertg:	building	consumption	NA	
	11362 (0.23%)	22607 (0.45%)	15755 (0.32%)	
	15767 (0.32%)	20771 (0.42%)	12998 (0.26%)	
petsg:	Yes	No	NA	
	1603 (0.03%)	3479 (0.07%)	44642 (0.90%)	
	16251 (0.33%)	5250 (0.11%)	28035 (0.57%)	

Continued

heatg:	CH	NCH	NA		
	24827 (0.50%)	14874 (0.30%)	10023 (0.20%)		
	17205 (0.35%)	20432 (0.41%)	11899 (0.24%)		
apcatg:	top	middle	low	below	NA
	4884 (0.10%)	33758 (0.68%)	1444 (0.03%)	188 (0.00%)	9450 (0.19%)
	4372 (0.09%)	34608 (0.70%)	2289 (0.05%)	389 (0.01%)	7878 (0.16%)
pcong:	First	Mt	Md	Inr	NA
	9013 (0.18%)	12680 (0.26%)	12731 (0.26%)	362 (0.01%)	14938 (0.30%)
	9409 (0.19%)	11535 (0.23%)	11289 (0.23%)	364 (0.01%)	16939 (0.34%)

3.2. Exploratory Data Analysis (EDA)

We can observe the superiority of the log-normal distribution over the normal distribution, as it provides a better fit to the rent_sqm data in Berlin for both 2015 and 2019, as shown in Figure 1 and Figure 2. This requires further statistical investigation.

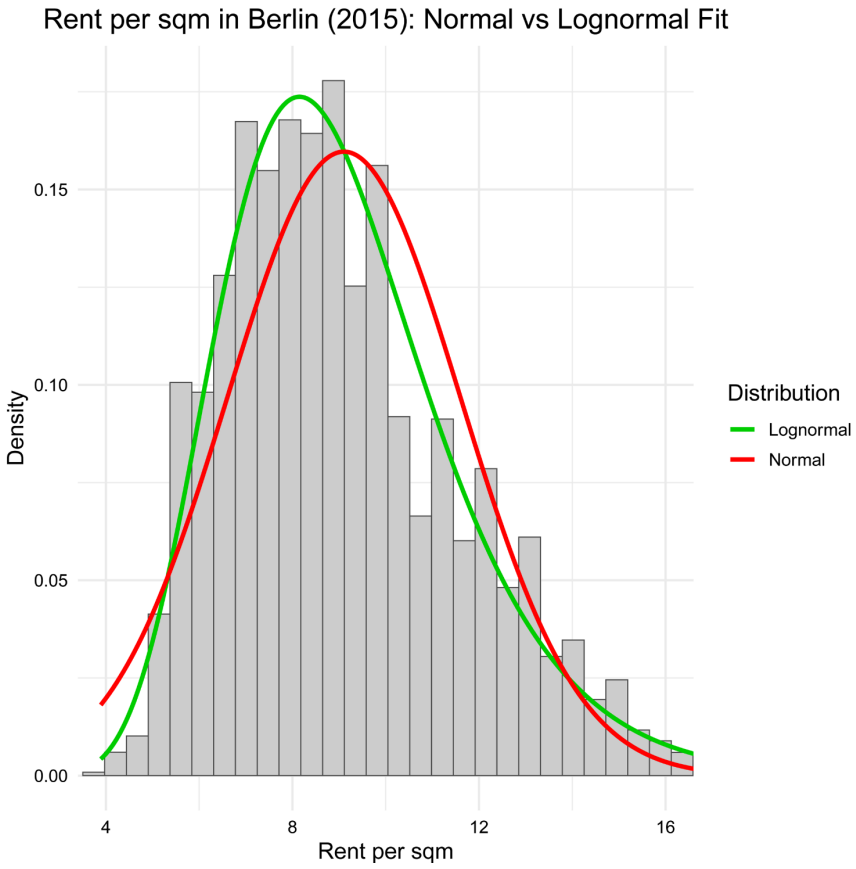


Figure 1. Normal and Log-Normal fittings of **rent_sqm for Berlin 2015.**

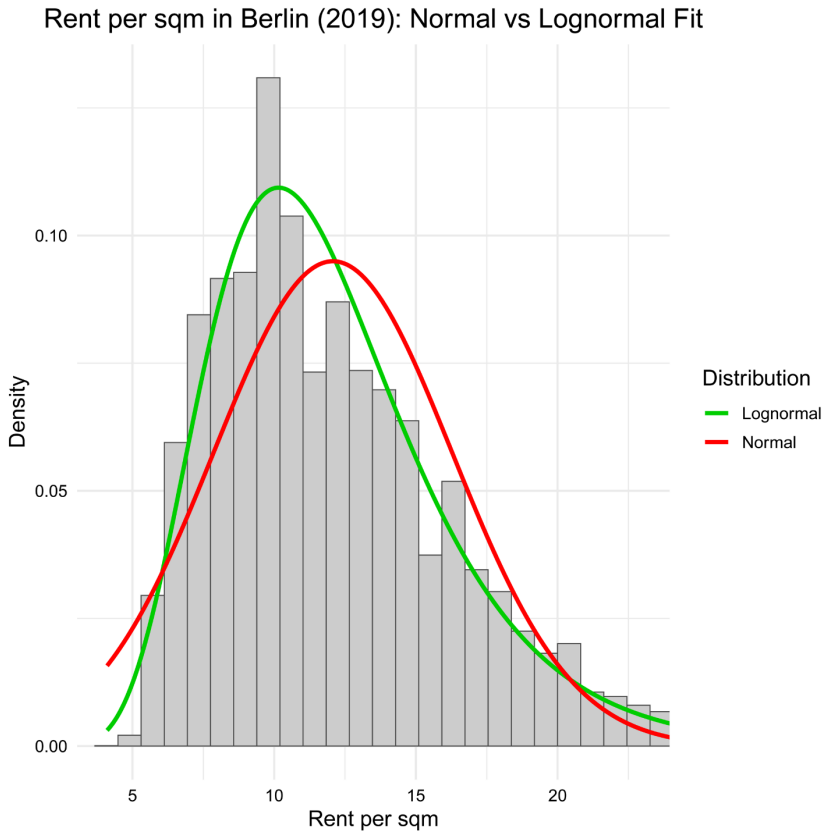


Figure 2. Normal and Log-Normal fittings of **rent_sqm for Berlin 2019.**

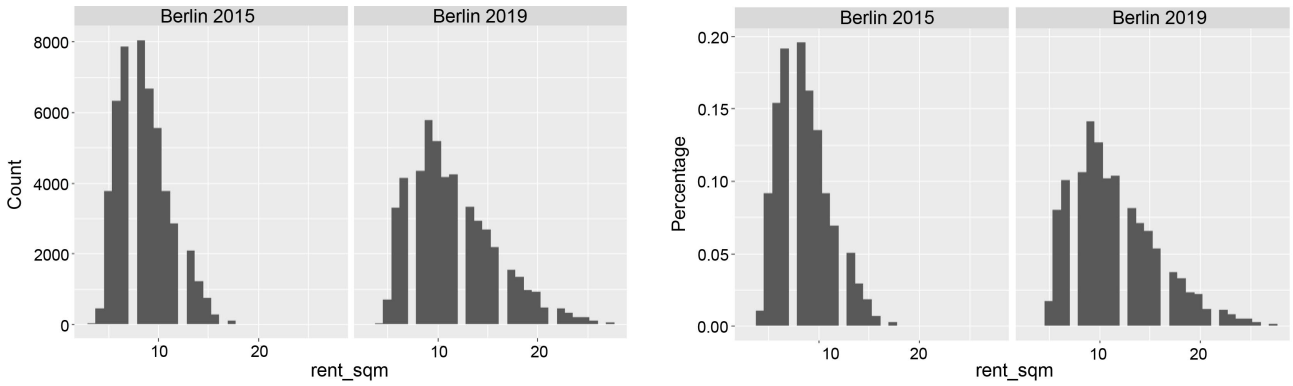
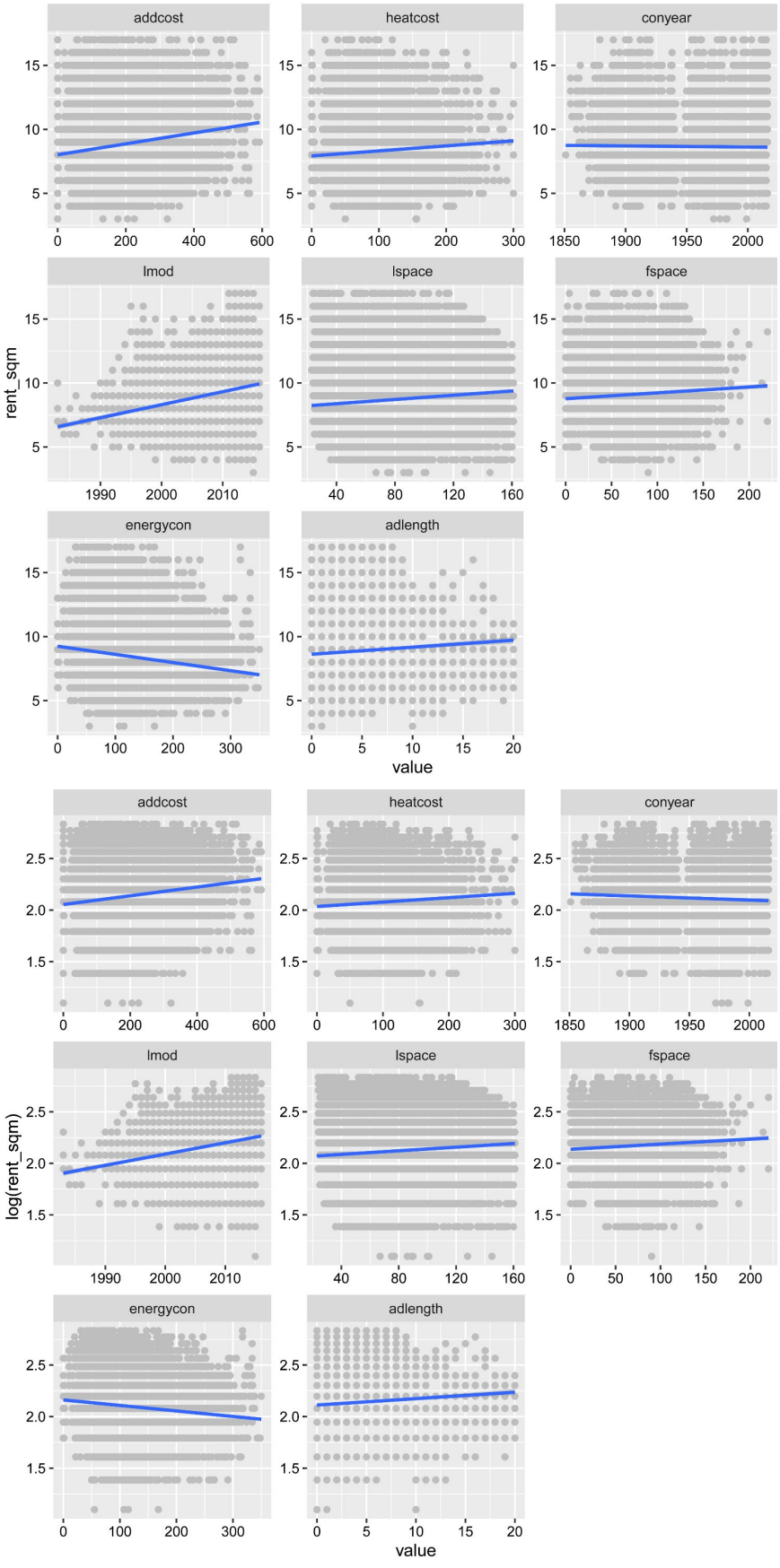


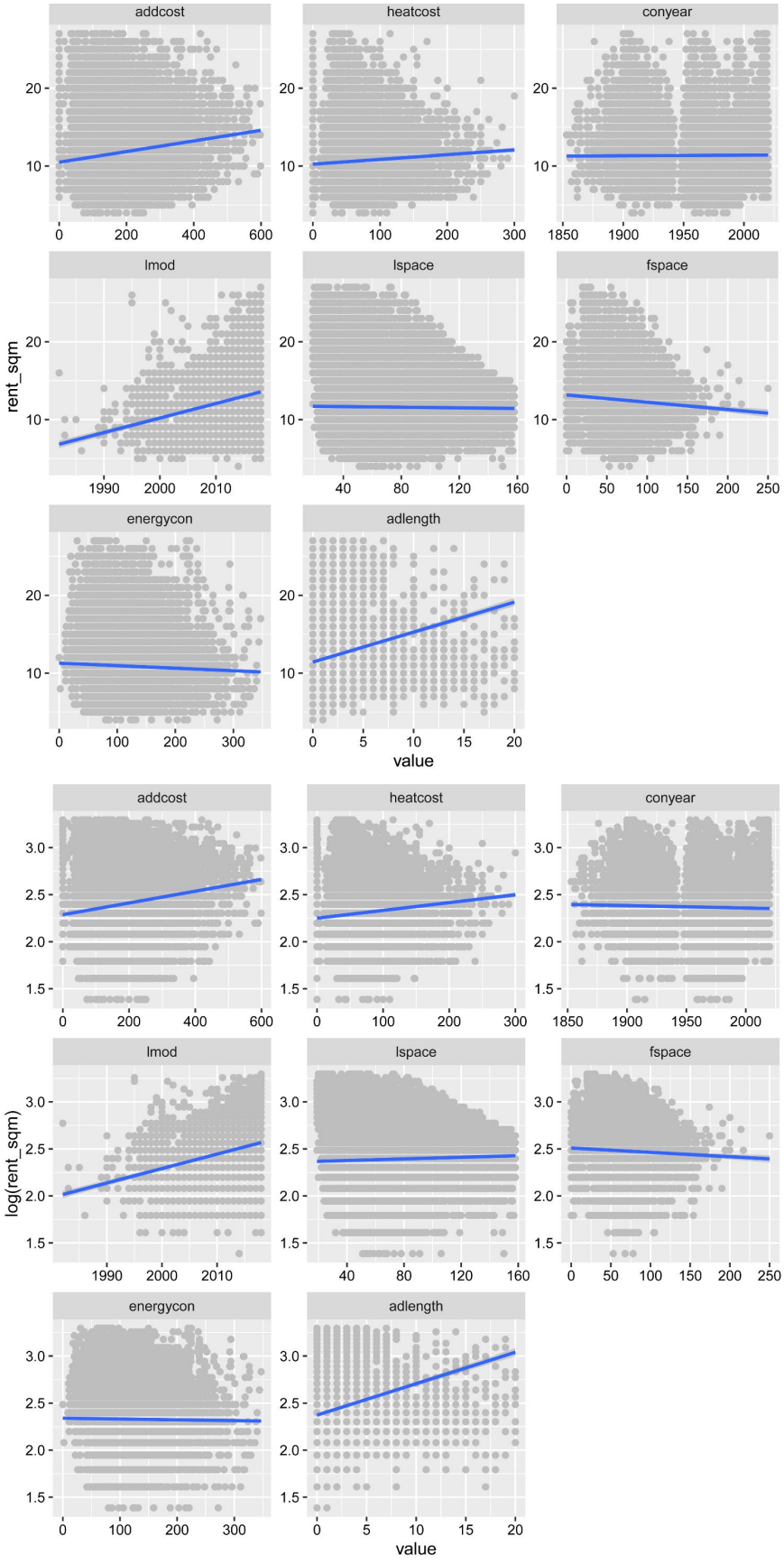
Figure 3. Histograms of response variable (**rent_sqm**): first column = **counts**, second column = **percentage.**

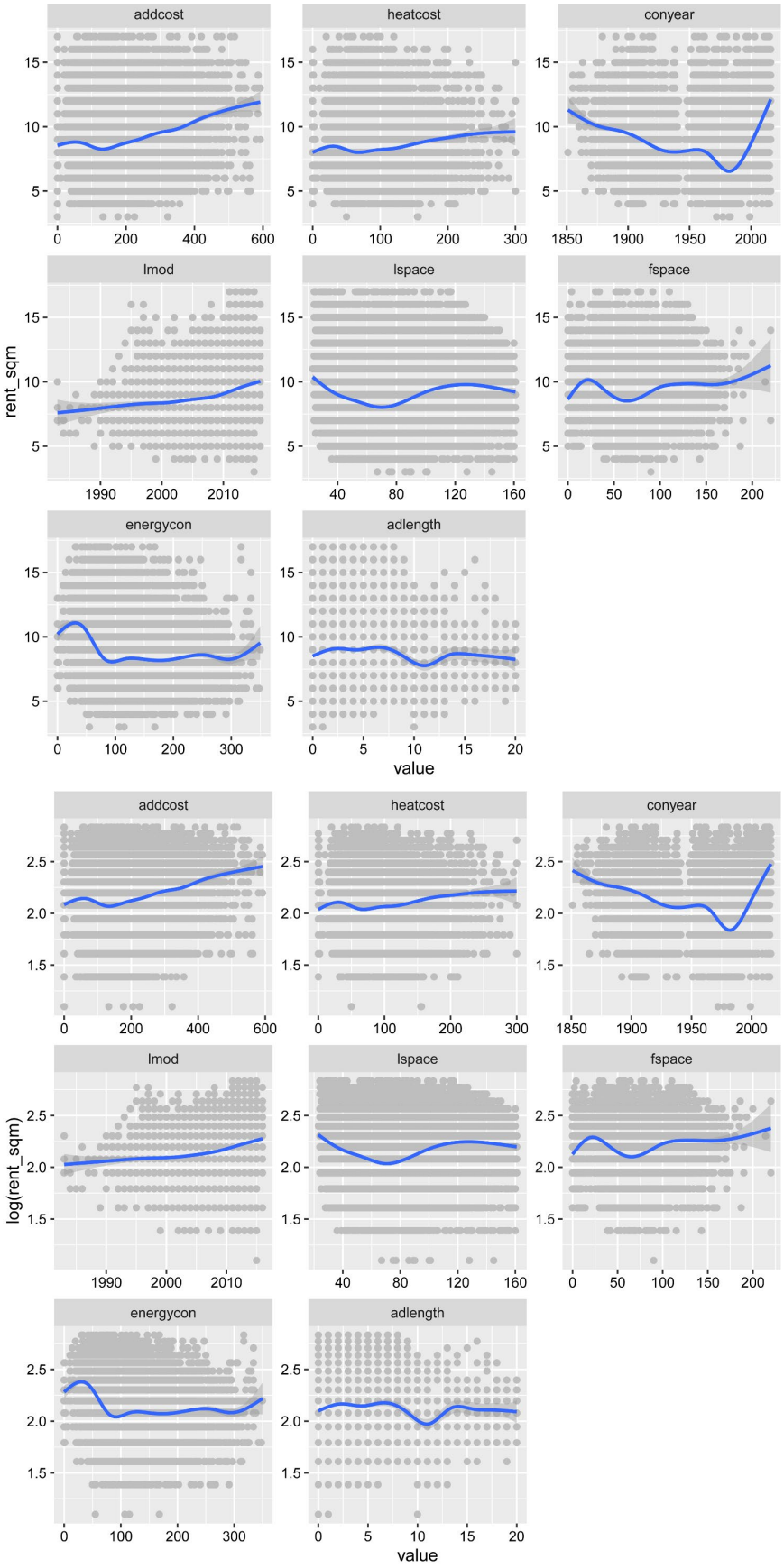
We also observe a significant shift in the histogram plots of rent_sqm for counts and percentages, as rent increased from below 20 in 2015 to above 20 in 2019.

3.3. Interpretation of Main Effects for the Quantitative and Qualitative Covariates

Looking at the above transformations on **rent_sqm** in Table 6, we may likely go with the log transformation for linear and non-linear covariates based on its suitability with respect to constant variance discussed in Section 2 and the effects of the covariates on **rent_sqm**.


addcost
heatcost
conyear
lmod
lspace
fspace
energycon
adlength
rent_sqm
log(rent_sqm)
value


addcost
heatcost
conyear
lmod
lspace
fspace
energycon
adlength
rent_sqm
log(rent_sqm)
value


addcost
heatcost
conyear
lmod
lspace
fspace
energycon
adlength
rent_sqm
log(rent_sqm)
value

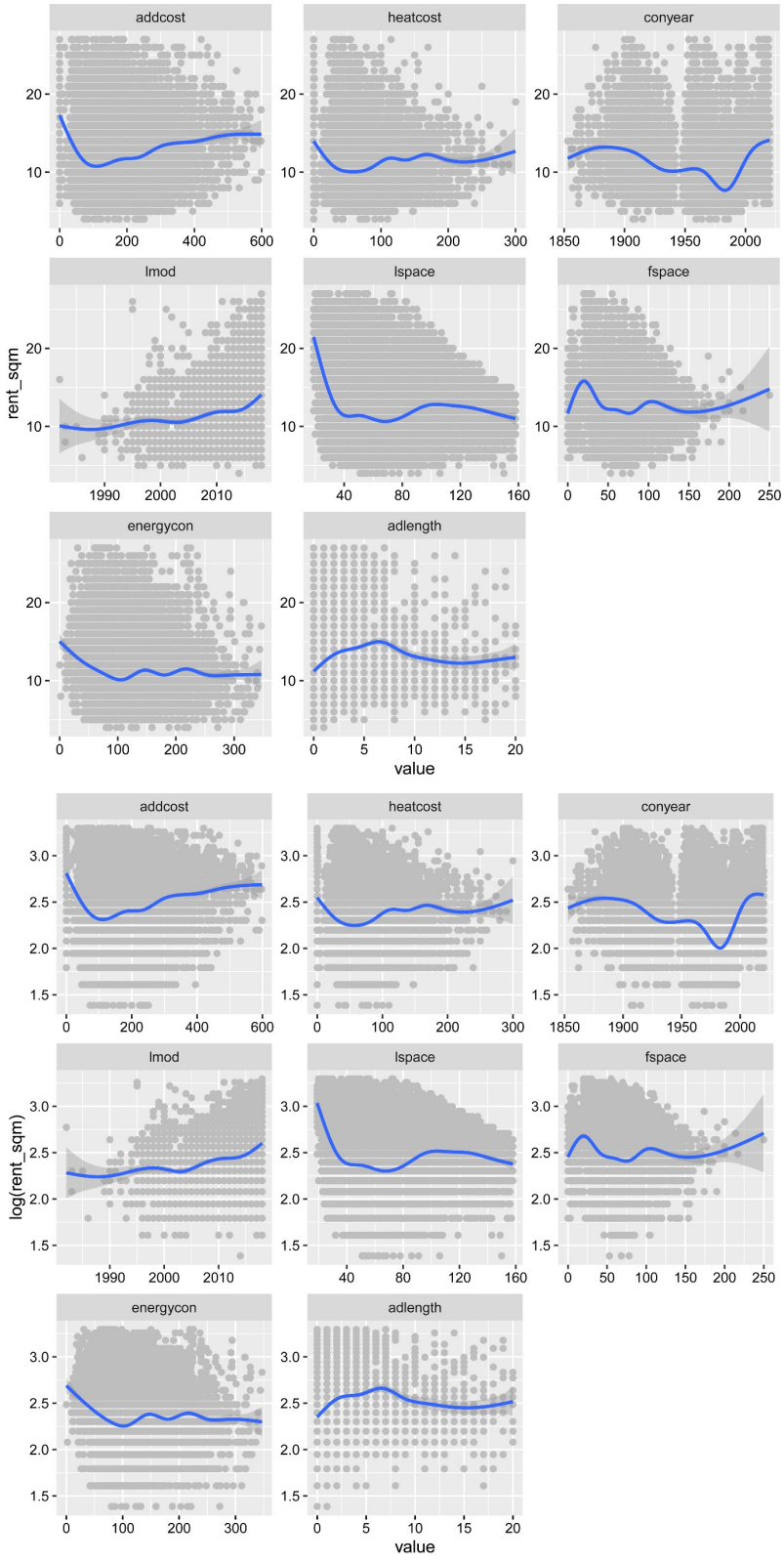


Figure 4. Scatter plots of quantitative covariates versus response (**rent_sqm**) with **linear smooth (LS)** and **non-linear smooth (NLS)**: first column = **rent_sqm** and second column = **log(rent_sqm)**. (First row) = **Berlin 2015 with LS**, (second row) = **Berlin 2019 with LS**, (third row) = **Berlin 2015 with NLS**, (fourth row) = **Berlin 2019 with NLS.**

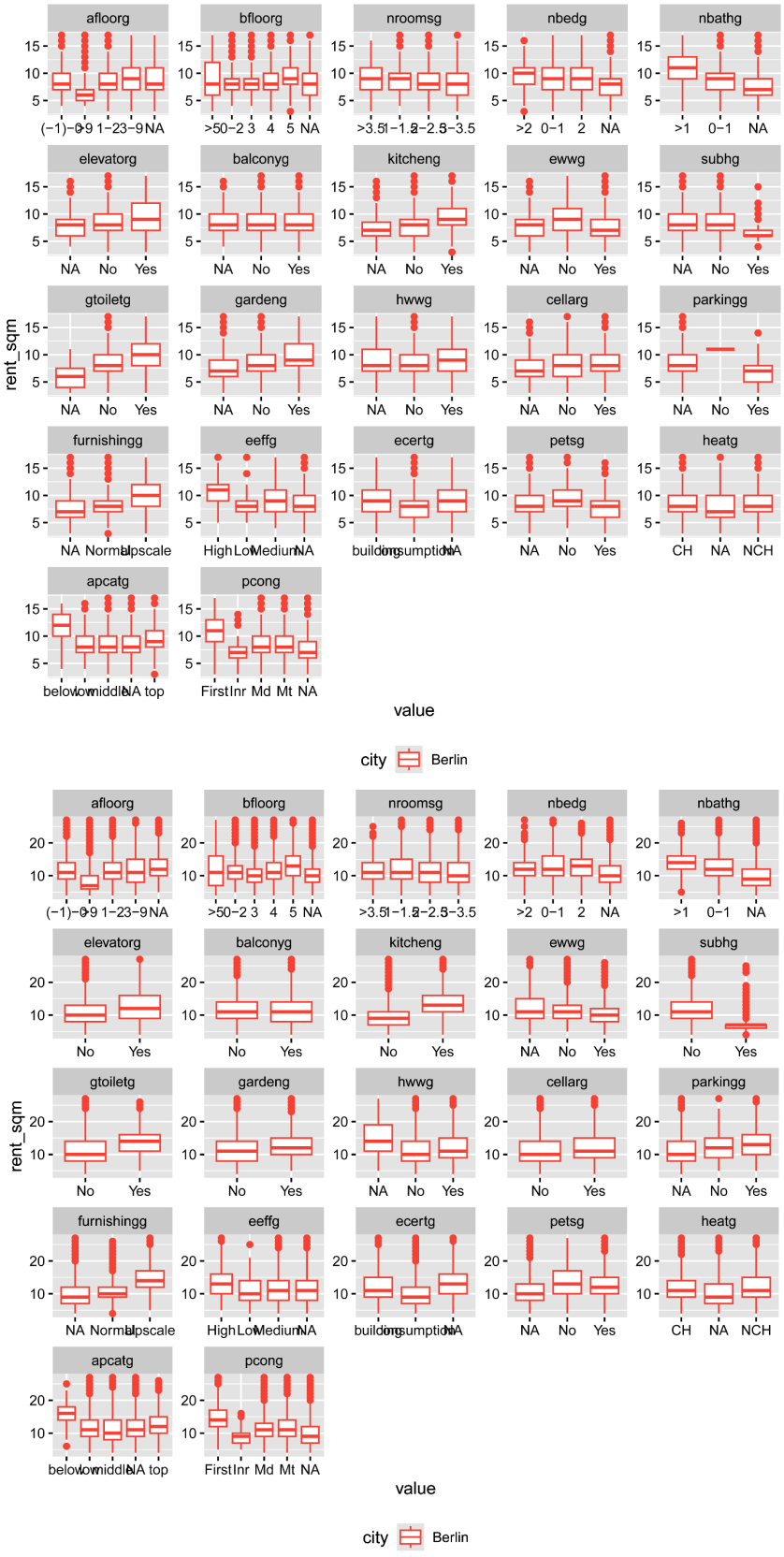


Figure 5. Box plots of qualitative covariates versus response (rent_sqm): first column = **Berlin 2015**, second column = **Berlin 2019.**

Table 6. Interpretation of main effects for the quantitative covariates on **rent_sqm** and **log(rent_sqm)** in Figure 4: first block = **Linear smooth**, second block = **Nonlinear smooth**.

Variables	Berlin 2015 rent_sqm	Berlin 2019 rent_sqm	Berlin 2015 log(rent_sqm)	Berlin 2019 log(rent_sqm)
addcost	Linear(increasing)	Linear (increasing)	Linear (increasing)	Linear (increasing)
heatcost	Linear (increasing)	Linear (increasing)	Linear (increasing)	Linear (increasing)
conyear	constant	constant	constant	constant
lmod	Linear (increasing)	Linear (increasing)	Linear (increasing)	Linear (increasing)
lspace	Linear (increasing)	constant	Linear (increasing)	nearly constant
fspace	Linear (increasing)	Linear (decreasing)	Linear (increasing)	constant
energycon	Linear (decreasing)	constant	Linear (decreasing)	constant
adlength	Linear (increasing)	Linear (increasing)	Linear (increasing)	Linear (increasing)
addcost	Quadratic	Quadratic	Quadratic	Quadratic
heatcost	nearly linear	Quadratic	nearly linear	nearly linear
conyear	Quadratic	cubic	Quadratic	nearly linear
lmod	Linear (increasing)	Quadratic	nearly constant	nearly constant
lspace	cubic	cubic	cubic	Linear (decreasing)
fspace	cubic	cubic	cubic	Quadratic
energycon	cubic	Quadratic	nearly constant	Quadratic
adlength	Quadratic	cubic	Quadratic	constant

Table 7. Interpretation for the qualitative covariates on **rent_sqm** in Figure 5.

Variables	Berlin 2015	Berlin 2019
afloorg	Yes	Yes
bfloorg	Yes	Yes
nroomsg	Yes	Yes
nbedg	Yes	No
nbathg	Yes	No
elevatorg	Yes	Yes
balconyg	No	No
kitcheng	Yes	Yes
ewwg	Yes	No
subhg	Yes	Yes
gtoiletg	Yes	Yes

Continued

gardeng	Yes	Yes
hwwg	Yes	Yes
cellarg	Yes	Yes
parkingg	Yes	Yes
furnishingg	Yes	Yes
eeffgg	Yes	Yes
ecertg	Yes	Yes
petsg	Yes	Yes
heatg	Yes	Yes
apcatg	Yes	Yes
pcong	Yes	Yes

4. Model Fittings and Predictions

We discuss how we select the type of model we use to fit the **rent_sqm** for Berlin rental properties in 2015 and 2019. To refine the regression model for the rent per square meter in Berlin, a stepwise backward regression was applied using the step() function in R. This method began with a full model containing all relevant predictors and iteratively removed nonsignificant variables based on the Akaike Information Criterion (AIC). The backward selection process ensured a more parsimonious model by retaining only the most influential variables, enhancing interpretability while maintaining predictive strength and minimizing model complexity. We fitted eight models ($M_1 \ldots M_8$) for the response variable, four each for the Berlin 2015 and 2019 datasets, as shown in the four cases with their respective summaries in Table 8.

Table 8. Model fitting summary with only main effect.

Berlin 2015	**case 1**	**case 2**	**case 3**	**case 4**
Adjusted R-square	0.3081	0.3255	0.3534	0.3645
Number of parameters (p)	42	42	50	52
Berlin 2019				
Adjusted R-square	0.3918	0.4218	0.353	0.4916
Number of parameters (p)	41	41	52	48

- **Case 1 (M_1 and M_2)**: Modeling **rent_sqm** with **linear covariates** for 2015 and 2019.
- **Case 2 (M_3 and M_4)**: Modeling **log(rent_sqm)** with **linear covariates** for 2015 and 2019.

- **Case 3 (M_5 and M_6)**: Modeling **rent_sqm** with **nonlinear covariates** for 2015 and 2019.
- **Case 4(M_7 and M_8)**: Modeling **log(rent_sqm)** with **nonlinear covariates** for 2015 and 2019

Based on the model fitting summary in Table 8, Case 4 was selected, as it best satisfied the model assumptions and achieved higher R^2 values across the datasets.

4.1. Model Fitting of log(rent_sqm) on Non-Linear Covariates for Berlin 2015 and Berlin 2019

Table 9. Berlin 2015.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−15.3309	1.6522	−9.28	0.0000
poly (addcost, 2)1	−0.0359	0.2912	−0.12	0.9018
poly (addcost, 2) 2	0.8549	0.2182	3.92	0.0001
poly (conyear, 2) 1	−2.4303	0.2424	−10.03	0.0000
poly (conyear, 2) 2	1.6659	0.2076	8.03	0.0000
lmod	0.0087	0.0008	10.61	0.0000
poly (lspace, 3) 1	−1.2691	0.5538	−2.29	0.0220
poly (lspace, 3) 2	2.0279	0.3350	6.05	0.0000
poly (lspace, 3) 3	−1.2145	0.2288	−5.31	0.0000
poly (fspace, 2) 1	−0.7175	0.3432	−2.09	0.0367
poly (fspace, 2) 2	−0.3325	0.3048	−1.09	0.2753
bfloorg 0 - 2	0.0061	0.0192	0.32	0.7517
bfloorg 3	−0.0014	0.0163	−0.09	0.9318
bfloorg 4	0.0490	0.0142	3.45	0.0006
bfloorg 5	0.0614	0.0137	4.48	0.0000
bfloorg NA	−0.0013	0.0238	−0.06	0.9559
nroomsg 1 - 1.5	0.0007	0.0284	0.03	0.9798
nroomsg 2 - 2.5	0.0507	0.0224	2.26	0.0237
nroomsg 3 - 3.5	0.0591	0.0176	3.35	0.0008
nbedg 0 - 1	−0.0556	0.0158	−3.52	0.0004
nbedg 2	−0.0482	0.0146	−3.31	0.0010
nbedg NA	−0.0689	0.0208	−3.31	0.0009
nbathg 0 - 1	−0.0138	0.0194	−0.71	0.4767
nbathg NA	−0.0626	0.0290	−2.15	0.0313

Continued

elevator No	−0.0414	0.0824	−0.50	0.6149
elevatorg Yes	0.0096	0.0823	0.12	0.9069
kitchen No	−0.0378	0.0832	−0.45	0.6493
kitcheng Yes	0.0315	0.0832	0.38	0.7051
ewwg No	0.0517	0.0338	1.53	0.1266
ewwg Yes	0.0683	0.0342	2.00	0.0455
subhg No	0.0096	0.0202	0.48	0.6339
subhg Yes	−0.2679	0.0482	−5.55	0.0000
gtoiletg Yes	0.0352	0.0152	2.32	0.0207
gardeng No	0.1277	0.0659	1.94	0.0527
gardeng Yes	0.1206	0.0662	1.82	0.0685
parking Yes	−0.0862	0.0512	−1.68	0.0922
furnishing Normal	0.0108	0.0147	0.73	0.4643
furnishingg Upscale	0.1274	0.0149	8.54	0.0000
eeffg Low	−0.0753	0.0478	−1.58	0.1152
eeffg Medium	−0.0580	0.0322	−1.80	0.0716
eeffg NA	−0.0189	0.0287	−0.66	0.5102
petsg No	−0.0377	0.0114	−3.30	0.0010
petsg Yes	−0.0479	0.0302	−1.58	0.1132
heatg NA	0.0267	0.0254	1.05	0.2946
heatg NCH	−0.0149	0.0087	−1.72	0.0856
apcatg low	−0.1461	0.0813	−1.80	0.0726
apcatg middle	−0.1150	0.0792	−1.45	0.1465
apcatg NA	−0.1284	0.0795	−1.61	0.1064
apcatg top	−0.0693	0.0795	−0.87	0.3834
pcong Inr	−0.1913	0.0615	−3.11	0.0019
pcong Md	−0.0606	0.0100	−6.04	0.0000
pcong Mt	−0.0291	0.0117	−2.48	0.0131
pcong NA	−0.0155	0.0205	−0.76	0.4485
Observations	2605			
R^2	0.377			
Adj. R^2	0.365			

Continued

Residual Std. Error	0.188 (df = 2552)
F Statistic	29.728*** (df = 52; 2552)
p-value	<2.2e−16

*p < 0.1; **p < 0.05; ***p < 0.01.

Table 10. Berlin 2019.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−13.2034	2.1819	−6.05	0.0000
poly (addcost, 2) 1	0.9758	0.3201	3.05	0.0023
poly (addcost, 2) 2	−0.2808	0.2557	−1.10	0.2724
heatcost	0.0009	0.0002	4.41	0.0000
poly (conyear, 2) 1	−1.2329	0.2958	−4.17	0.0000
poly (conyear, 2) 2	2.7830	0.2552	10.90	0.0000
lmod	0.0077	0.0011	7.10	0.0000
poly (lspace, 3) 1	−2.7888	0.5671	−4.92	0.0000
poly (lspace, 3) 2	1.4259	0.3374	4.23	0.0000
poly (lspace, 3) 3	−1.3228	0.2577	−5.13	0.0000
poly (energycon, 2) 1	0.3285	0.2503	1.31	0.1896
poly (energycon, 2) 2	0.5836	0.2260	2.58	0.0099
poly (adlength, 3) 1	0.0461	0.2257	0.20	0.8383
poly (adlength, 3) 2	−0.7181	0.2263	−3.17	0.0015
poly (adlength, 3) 3	0.4056	0.2166	1.87	0.0613
afloorg >9	−0.0672	0.0668	−1.01	0.3143
afloorg 1 - 2	0.0059	0.0230	0.26	0.7982
afloorg 3 - 9	0.0577	0.0237	2.43	0.0151
afloorg NA	−0.1273	0.0483	−2.64	0.0085
bfloorg 0 - 2	0.0576	0.0322	1.79	0.0742
bfloorg 3	0.0056	0.0288	0.20	0.8448
bfloorg 4	0.0649	0.0242	2.68	0.0075
bfloorg 5	0.1216	0.0215	5.65	0.0000
bfloorg NA	0.1357	0.0413	3.28	0.0011
nroomsg 1 - 1.5	−0.0444	0.0449	−0.99	0.3231

Continued

nroomsg 2 - 2.5	0.0550	0.0350	1.57	0.1159
nroomsg 3 - 3.5	0.0192	0.0281	0.68	0.4942
elevatorg Yes	0.0900	0.0168	5.36	0.0000
kitcheng Yes	0.1231	0.0139	8.84	0.0000
ewwg No	−0.0580	0.0175	−3.32	0.0009
ewwg Yes	−0.0091	0.0173	−0.53	0.5972
subhg Yes	−0.4174	0.1078	−3.87	0.0001
cellarg Yes	−0.0260	0.0146	−1.78	0.0749
parkingg No	0.0970	0.0738	1.31	0.1890
parkingg Yes	−0.0655	0.0163	−4.02	0.0001
furnishing Normal	−0.0390	0.0295	−1.32	0.1872
furnishingg Upscale	0.1026	0.0295	3.47	0.0005
petsg No	−0.0016	0.0205	−0.08	0.9369
petsg Yes	−0.0444	0.0149	−2.97	0.0030
heatg NA	0.0011	0.0530	0.02	0.9833
heatg NCH	−0.0573	0.0144	−3.98	0.0001
apcatg low	−0.0641	0.0924	−0.69	0.4875
apcatg middle	−0.0268	0.0839	−0.32	0.7494
apcatg NA	−0.0051	0.0856	−0.06	0.9521
apcatg top	0.0280	0.0839	0.33	0.7386
pcong Inr	−0.1959	0.0850	−2.31	0.0213
pcong Md	−0.0292	0.0210	−1.39	0.1650
pcong Mt	−0.0562	0.0223	−2.52	0.0119
pcong NA	0.0227	0.0381	0.59	0.5522
Observations	1231			
R^2	0.511			
Adj. R^2	0.492			
Residual Std. Error	0.211 (df = 1182)			
F Statistic	25.778^{***} (df = 48; 1182)			
p-value	<2.2e−16			

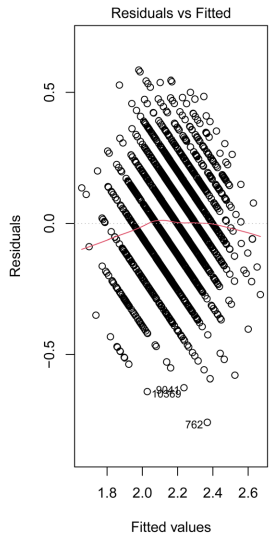
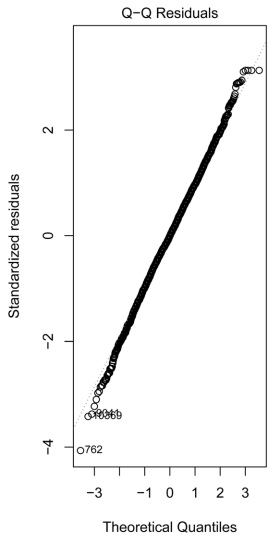
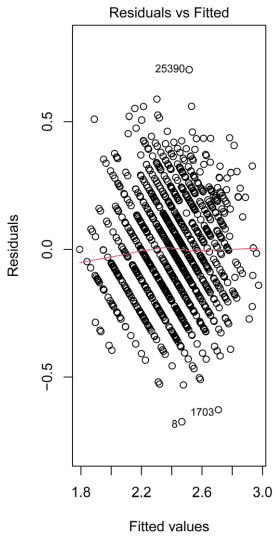
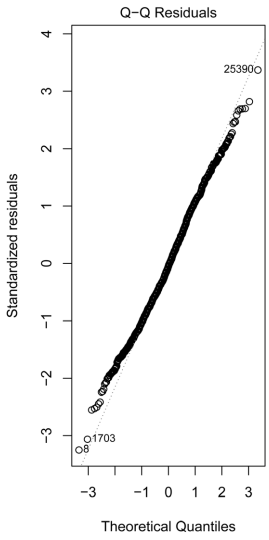
*p < 0.1; **p < 0.05; ***p < 0.01.

4.2. Residual Plots of Model Fittings

We plot the residuals versus the fitted values to check for a trend and assess the plausibility of the linearity assumption discussed in Section 2. Additionally, we plot the QQ plots of the covariates versus the theoretical normal quantiles to check if they form a straight line, which helps assess the plausibility of the normality assumption discussed in Section 2.

From the plots in Table 11, we find that the fitted models do not relatively violate the linear regression assumptions in Section 2.

Table 11. Residual plots of model fittings for **Berlin 2015** and **Berlin 2019**.

city	Berlin 2015		Berlin 2019	
				

4.3. Model Predictions of rent_sqm for the Main Effect Models

In this section, we predict **rent_sqm** using the main effect models in Table 9 and Table 10, based on influential variables identified through pairwise selection (Table 12 and Table 13). Predictions use the **median** for continuous covariates and the **mode** for qualitative covariates. For each variable of interest, values are generated across the 5th to 95th percentiles while holding other covariates at their median or modal values. All adjusted $GVIF^{1/(2\cdot Df)}$ values were below 2, indicating no multicollinearity concerns; therefore, all predictors were retained, confirming the model's statistical robustness.

We observe that the model predictions in Table 12, which illustrate the major influential quantitative covariates selected through the AIC procedure, exhibit predominantly nonlinear relationships with rent_sqm in both years (2015 and 2019), apart from last modernization, which consistently showed a linear trend. Also, in 2019, variables such as additional costs (addcost), heat cost (heatcost), and energy condition (energycon) show increasing trends, while living space (lspace), advertisement length (adlenght), and the year in which the object was

built (conyear) show a decreasing pattern. These effects, derived from model predictions, suggest that major numerical factors influence rent in complex, nonlinear ways rather than through simple linear associations.

Table 12. Model predictions of **rent_sqm** for the influential **quantitative covariates**.

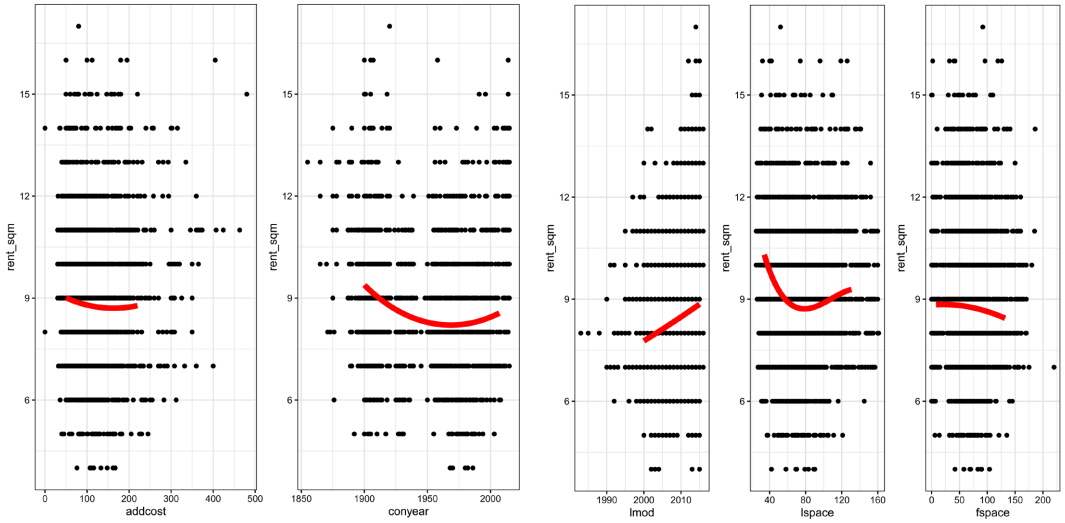
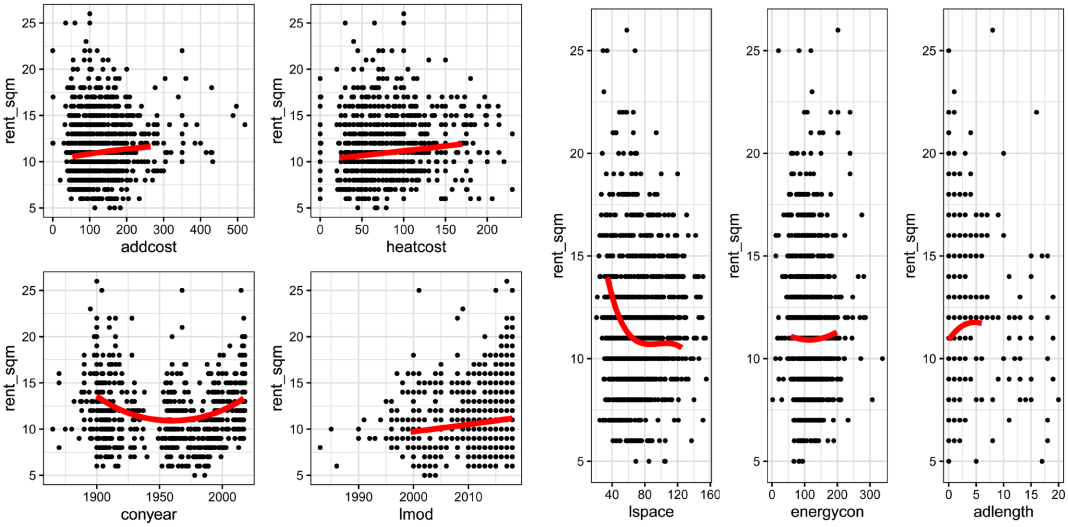
Berlin 2015	
Berlin 2019	

Table 13. Model predictions of **rent_sqm** for the influential **qualitative covariates**.

Variables	Categories	2015	2019
afloorg	(−1) - 0		10.30
	1 - 2		10.36
	3 - 9		10.91
	>9		9.63
	NA		9.07

Continued

bfloorg	0 - 2	8.41	10.83
	3	8.34	10.29
	4	8.78	10.91
	5	8.89	11.55
	>5	8.36	10.23
	NA	8.35	11.72
nroomsg	1 - 1.5	8.35	9.88
	2 - 2.5	8.78	10.91
	3 - 3.5	8.85	10.53
	>3.5	8.34	10.33
nbedg	0 - 1	8.78	
	2	8.84	
	>2	9.28	
	NA	8.66	
nbathg	0 - 1	8.78	
	>1	8.90	
	NA	8.66	
elevatorg	Yes	9.24	11.94
	No	8.78	10.91
	NA	9.15	
kitcheng	Yes	9.41	10.91
	No	8.78	9.65
	NA	9.11	
ewwg	Yes	8.92	10.81
	No	8.78	10.30
	NA	8.33	10.91
subhg	Yes	6.65	7.19
	No	8.78	10.91
	NA	8.69	
parkingg	Yes	8.05	10.22
	No		12.03
	NA	8.78	10.91

Continued

	Upscale	8.78	10.91
furnishingg	Normal	7.81	9.47
	NA	7.73	9.85
	High	8.94	
	Medium	8.44	
eeffgg	Low	8.30	
	NA	8.78	
	Yes	8.37	10.91
petsg	No	8.45	11.39
	NA	8.78	11.41
	Top	9.19	11.53
	Middle	8.78	10.91
apcatg	Low	8.51	10.51
	Below	9.85	11.21
	NA	8.67	11.15

5. Results

5.1. Summary of Results

We observe in Figure 1 and Figure 2 that in 2015, the lognormal curve provided a superior fit to the empirical data, particularly around the mode and in the upper tail. The histogram revealed an apparent right skew, which the lognormal distribution captured effectively. The normal distribution, by contrast, imposed symmetry on an asymmetric dataset, resulting in visible deviation from the observed pattern, especially at the extremes. In 2019, this pattern persisted. The rental data maintained a skewed distribution, and the lognormal curve once again fit the distribution more accurately, capturing both the peak and the tail behavior. The normal curve underestimated the density at lower values and overestimated it in the central range, confirming its misalignment.

In addition, in Figure 3, a significant shift is observed in the histogram plots of **rent_sqm** between Berlin 2015 and 2019. For example, in Munich 2015, most values of **rent_sqm** were below 20 Euros, whereas in 2019, they exceeded 20 Euros. This indicates that the rent price increases over time, which is also confirmed by our prediction. For example, the **predicted rent_sqm** increased in Berlin from 2015 to 2019 by 15.94%, 24.26%, 20.6%, 24.15%, and 30.22%, with apartments that have a kitchen, Upscale furnishing, First occupancy condition, central heating, and do not allow pets, respectively.

From Table 12, the influence of quantitative covariates on the **predicted**

rent_sqm is summarized below.

- For **Berlin 2015**, each variable exhibits an effect on the **rent_sqm**, as observed by their non-constant predicted lines. The variable, last modernization, enters the model linearly, showing a positive trend with **rent_sqm**, whereas the remaining variables demonstrate nonlinear relationships.
- Also, for **Berlin 2019**, all variables impact the **rent_sqm**. The additional cost, heat cost, and last modernization variables entered linearly, each displaying an increasing trend with **rent_sqm**, while the other variables enter the model in a nonlinear form.

In Table 13, we can also summarise the pattern of the **predicted rent_sqm** for the influential qualitative covariates as follows.

- The predicted **rent_sqm** is **highest in Berlin**, estimated at 12.03 euros in 2019.
- **afloor:** In Berlin 2019, the predicted **rent_sqm** is at the lowest (9.63 Euros) with apartment floor >9. It increased with apartment floor (−1) - 0 (10.30 Euros), followed by apartment floor 1- 2 (10.36 Euros) and it is at the highest (10.91 Euros) with apartment floor 3 - 9 (the mode).
- **building floors (bfloor):** With five building floors apartments, our predicted **rent_sqm** is at the highest (8.89 and 11.55 Euros) for Berlin 2015, Berlin 2019, but is at the lowest (8.34 and 10.29 Euros) with three building floors apartments.
- **number of rooms (nrooms):** In **Berlin 2015**, our **predicted rent_sqm** value is at the highest (8.85 Euros) with apartments that have **the number of rooms 3 - 3.5** while in **Berlin 2019**, the apartments with **the number of rooms 2 - 2.5** has the highest **predicted rent_sqm** (10.91 Euros). On the other hand, the predicted **rent_sqm** is at the lowest with apartments that have 1-1.5 (8.35, 9.88, and 12.96 Euros) for Berlin 2015, Berlin 2019.
- **number of bedrooms (nbed)**: In **Berlin 2015**, we can see **an increasing trend** in the **predicted rent_sqm** (8.78, 8.84, and 9.28 Euros) with respect to the order of the categories of the **number of bedrooms (0 - 1, 2, and** >2**)** of an apartment. Thus, rent_sqm seems to increase in Berlin for apartments with a higher number of bedrooms.
- **number of bathrooms (nbath):** In Berlin 2015, we can also see **an increasing trend** in the **predicted rent_sqm** (8.78 and 8.90 Euros) with respect to the order of the categories of the **number of bathrooms (0 - 1, and** >1**)** of an apartment. Thus, rent_sqm seems to increase with apartments that have a higher number of bathrooms in Berlin (and vice versa).
- **elevator:** We can see **an increase** in the **predicted rent_sqm** (9.24 and 11.94 Euros) for apartments with an elevator in Berlin 2015 and Berlin 2019, respectively unlike the apartments without an elevator where the **predicted rent_sqm** are respectively (8.78 and 10.91 Euros). Thus, rent_sqm seems to increase with apartments that have an elevator (and vice versa).
- **kitchen:** We can see **an increase** in the **predicted rent_sqm** (9.41 and 10.91 Euros) for apartments with a kitchen in Berlin 2015 and Berlin 2019, respectively unlike the apartments without a kitchen where the **predicted rent_sqm** are

respectively (8.78 and 9.65 Euros). Thus rent_sqm seems to increase with apartments that have a kitchen (and vice versa).

- **eww:** The **predicted rent_sqm** is higher with apartments that have the inclusion of warm water consumption in the energy consumption value calculation (8.92 and 10.81 Euros) in both Berlin 2015 and 2019 respectively, compared to the apartments that do not have it (8.78 and 10.30 euros). Thus rent_sqm seems to increase for apartments that have the inclusion of warm water consumption in the energy consumption value calculation in Berlin (and vice versa).
- **subh:** The **predicted rent_sqm** is lower with apartments that have a certificate of eligibility for public housing (6.65 and 7.19 Euros) in both Berlin 2015 and 2019, respectively, compared to the apartments that do not have it (8.78 and 10.91 Euros). Thus rent_sqm seems to increase with apartments that do not have a certificate of eligibility for public housing in Berlin (and vice versa).
- **gtoilet:** The **predicted rent_sqm** is higher with apartments that have a guest toilet (9.09 Euros) in Berlin 2015, compared to the apartments with no guest toilet (8.78 Euros). Thus, rent_sqm seems to increase with apartments that have a guest toilet in Berlin (and vice versa).
- **garden:** With apartments that have a garden in Berlin 2015, the **predicted rent_sqm** is lower (8.71 Euros) compared to apartments that do not have a garden (8.78 Euros). Thus, rent_sqm seems to decrease with apartments that have a cellar in Berlin (and vice versa)
- **cellar:** With apartments that have a cellar in Berlin 2019, the **predicted rent_sqm** is lower (10.91 Euros) compared to apartments that do not have cellars (11.20 Euros). Thus, rent_sqm seems to decrease with apartments that have a cellar in Berlin (and vice versa)
- **parking space:** In Berlin 2019, the **predicted rent_sqm** is higher (10.22 Euros) with apartments that have a parking space, compared to apartments that do not have a parking space (12.03 Euros). Thus rent_sqm seems to increase in Berlin with apartments that have a parking space (and vice versa).
- **furnishing:** The predicted **rent_sqm** is at the highest with apartments that have **Upscale furnishing** for **Berlin 2015** and **Berlin 2019**. Also, with Upscale furnishing apartments, the predicted **rent_sqm** increased from 2015 to 2019 by 24.26%. It equally increased from Normal to Upscale furnishing apartments by 12.42% and 15.21%, for Berlin 2015 and 2019, respectively. Thus, rent_sqm seems to increase with apartments that have Upscale furnishing in Berlin (and vice versa), as well as with respect to time.
- **energy efficiency rating (eeff)**: We can also see **an increasing trend** in the **predicted rent_sqm** with respect to the order of the categories of **energy efficiency rating (Low, Medium, and High)** (8.30, 8.44, and 8.94 Euros) in **Berlin 2015**. Thus rent_sqm seems to increase in Berlin with respect to the order of energy efficiency rating categories (Low, Medium, and High) (and vice versa).
- **pets:** The **predicted rent_sqm** is lower with apartments that allow pets (8.37

and 10.91 Euros) in Berlin 2015 and 2019, compared to apartments that do not allow pets (8.45 and 11.39 Euros), thereby decreasing by 23.28% and 25.28% for Berlin 2015 and 2019, respectively. Thus, rent_sqm seems to decrease with apartments that allow pets in Berlin (and vice versa)

- **heat:** Our predicted **rent_sqm** is higher with apartments that make use of the **central heating (CH)** as their heating type (8.78 and 10.91 Euros) for Berlin 2015 and 2019, respectively, compared to the apartments that make use of the **non-central heating (NCH)** as their heating type (8.65 and 10.31 Euros). Thus, rent_sqm appears to increase with apartments that utilize central heating as their primary heating type in Berlin (and vice versa).
- **apartment categories (apcat)**: Our predicted **rent_sqm** is at the highest with the **below category** apartments (9.85 Euros) for **Berlin 2015** while for **Berlin 2019**, it is at the highest with the **top category** apartments (11.53 Euros). On the other hand, our predicted **rent_sqm** is at its lowest for the **low category** apartments (8.51 and 10.51 Euros) in **Berlin 2015 and 2019**, respectively. Thus, rent_sqm seems to increase with low and top-category apartments in Berlin (and vice versa).
- **property condition categories (pcon)**: Our predicted **rent_sqm** is at the highest with the **First occupancy condition** apartments (9.32 and 11.24 Euros) for **Berlin 2015** and **2019**, respectively, thereby increasing by 20.0% from 2015 to 2019. Thus, rent_sqm seems to increase with the first occupancy condition apartments in Berlin compared to other apartment condition categories.

5.2. Discussion on Research Questions

5.2.1. RQ1: Does a Relationship Exist between the Response Variable (rent_sqm) and the Selected Predictor Variables?

The analysis indicates strong associations between rent_sqm and most predictors. In both the Berlin 2015 and 2019 datasets, all examined variables influenced rent_sqm as shown in Figure 4 and Figure 5, Table 6, Table 7, Table 12, and Table 13. For example, the last modernization variable showed a consistent linear relationship with rent_sqm, while other quantitative variables exhibited nonlinear trends. Also, the boxplots in Figure 5 as interpreted in Table 7, provide visual evidence of the relationship between rent_sqm and the qualitative covariates. Observable patterns and non-parallel trends across category levels suggest that these variables influence rental behavior. These relationships validate the importance of diverse housing features in determining rent prices. This finding is consistent with housing studies that emphasize the role of structural and locational features in determining prices [4] [5].

Moreover, the results align with broader market trends. According to JLL's Housing Market Overview, Berlin's 2024 housing market is expected to face growing demand, increased immigration, and lagging construction. Despite rent price stabilization and regulatory efforts, declining completions and affordability challenges signal persistent pressure and possible structural limits in supply, investment, and population-driven housing demand. This suggests that various

factors, including those studied, contribute to rent variations [1].

5.2.2. RQ2: Is a Transformation of Response Variable (rent_sqm) Necessary to Meet the Assumptions of Linear Regression?

Yes, first, density plots were generated for the years 2015 and 2019 to assess the appropriate distributional model for rent per square meter in Berlin. Each histogram was overlaid with both standard (red) and lognormal (green) fitted density curves, as shown in **Figure 1** and **Figure 2**. Across both years, the lognormal distribution consistently exhibited a closer approximation to the observed rent data. Secondly, a log transformation was applied to rent_sqm to address nonnormality and heteroskedasticity. This transformation statistically confirmed the superiority of lognormal over the normal distribution as observed in **Figure 1** and **Figure 2**, by improving the model's explanatory power, as indicated by the increased adjusted R-squared values in case 4 of **Table 8** (e.g., 0.3645 for 2015 and 0.4916 for 2019), and a better alignment with regression assumptions. This transformation approach also supports previous findings and aligns with econometric methods often used in real estate analyses, where log-linear models account for skewed distributions and non-constant variance [4] [5] [28] [29].

5.2.3. RQ3: Which Predictor Variables Significantly Affect the Rental Price Per Square Meter in Berlin's Housing Market?

The study identified several influential predictors impacting rental prices:

- **Quantitative Covariates:** Last modernization year (linear), additional cost (linear), and heat cost (linear) significantly influenced rent_sqm. Other variables like living space and construction year significantly entered the model nonlinearly, showing complex effects across the two years 2015 and 2019 as shown in **Table 12**.
- **Qualitative Covariates:** Upscale furnishing increased rent by over 24%, while the presence of amenities (e.g., elevator, kitchen, guest toilet, high energy efficiency ratings, the inclusion of warm water consumption in the energy consumption value calculation) correlated with higher rents as shown in **Table 13**. Conversely, pets allowed was correlated with lower rents. These results are consistent with the existing literature, which highlights the importance of property characteristics and amenities in determining rental values [5] [30].

5.3. Contribution of the Study

This study contributes to six important domains:

- **Contribution to Literature:** This study advances statistical modeling literature by integrating lognormal regression with nonlinear covariate transformations to better capture skewed rental price distributions. The approach improves model flexibility and predictive accuracy, offering a refined methodological framework for urban housing analysis and policy applications [30].
- **Statistical Applications:** The study demonstrates the effective use of advanced regression techniques, including log transformation, nonlinear covariates, and

GVIF diagnostics, in analyzing large urban rental datasets. This application provides a practical teaching resource for applied statistics and data science courses [21] [22].

- **Educational Leadership and Evidence-Based Policy:** By identifying key cost drivers affecting rental affordability, the study supports data-informed decision-making in higher education, particularly in addressing student and faculty housing equity challenges [8].
- **Cross-National Implications:** Berlin's rental market dynamics mirror those of other global academic cities, allowing the findings to inform housing strategies across international contexts with both market-driven and regulated systems [6].
- **Policy Innovation and Housing Equity:** The analysis highlights the role of modernization and energy efficiency in shaping affordability, contributing insights for sustainable housing policies that balance market performance with equity considerations [7].
- **Institutional Housing Strategy Improvement:** The study offers a replicable, empirically grounded framework to guide institutional housing planning, negotiations, and subsidy design. By identifying major rent drivers, it supports targeted affordability interventions in high-demand cities in Germany, the United States, and beyond [9] [31].

6. Conclusions

This study applied multiple linear regression with nonlinear covariates to model rental prices per square meter in Berlin's housing market using a combined qualitative and quantitative dataset. The results indicate that rent levels are significantly influenced by furnishing quality, year of last modernization, floor level, and apartment category, with both linear and nonlinear effects, reinforcing the need for flexible statistical approaches in housing analysis [4] [5]. Methodologically, the study extends existing research by incorporating nonlinear covariates into lognormal regression models. Among the tested specifications, the log-transformed model with nonlinear terms outperformed others for both the 2015 and 2019 datasets, achieving adjusted R^2 values of 0.3645 and 0.4916, respectively. These findings confirm the suitability of logarithmic transformation for handling skewness and heteroscedasticity in rental data, consistent with prior literature [4] [5] [28] [30] [32]. Diagnostic measures further supported the statistical validity and interpretability of the model. From a practical perspective, the findings offer guidance for real estate professionals, urban planners, and institutional leaders by highlighting the importance of modernization, energy efficiency, and structural features in rental policy design, particularly in high-demand urban markets [1] [2] [32]. For higher education institutions, the study offers a replicable framework to inform housing negotiations, subsidy allocation, and planning in student-dense cities, aligning with broader calls for data-driven policy in housing equity [5] [7] [9]. Future research may extend this framework through time-series forecasting,

spatial econometric methods, or comparative analyses across major European and U.S. cities.

Finally, the authors recommend integrating major housing cost drivers and predictive modeling into higher education housing and affordability decisions. The study is limited to Berlin data from 2015 and 2019, which restricts generalizability and excludes post-pandemic changes.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Jones Lang LaSalle (JLL) (2024) Housing Market Overview—H2 2024. https://www.jll.de/en/trends-and-insights/research/housing-market-overview

[2] Lutz, E. (2020) The Housing Crisis as a Problem of Intergenerational Justice: The Case of Germany. *Intergenerational Justice Review*, **1**, 2020.

[3] Meyberg, C., Rendtel, U. and Leerhoff, H. (2024) Flat Rent Price Prediction in Berlin with Web Scraping. *AStA Wirtschafts- und Sozialstatistisches Archiv*, **18**, 245-278. https://doi.org/10.1007/s11943-024-00340-6

[4] Malpezzi, S., *et al*. (2003) Hedonic Pricing Models: A Selective and Applied Review. *Housing Economics and Public Policy*, **1**, 67-89.

[5] Yoshida, T., Murakami, D. and Seya, H. (2024) Spatial Prediction of Apartment Rent Using Regression-Based and Machine Learning-Based Approaches with a Large Dataset. *The Journal of Real Estate Finance and Economics*, **69**, 1-28. https://doi.org/10.1007/s11146-022-09929-6

[6] Brookings Institution (2023) How a University-Community Homesharing Collective is Creating a New Model for Affordable Housing in West Philadelphia.

[7] Pivo, G. (2022) Green Buildings and Rental Premiums: A Meta-Analysis. *Journal of Sustainable Real Estate*, **14**, 1-16.

[8] Fullan, M. (2020) Leading in a Culture of Change. 2 Edition, John Wiley & Sons, Incorporated.

[9] German Academic Exchange Service (DAAD) (2024) Internationalisation Only Successful with Sufficient Living Space for Students. Press Release.

[10] Czado and Brechmann (2021) Lecture Slides on GLM, Study Material from the Research Group Mathematical Statistics in the Department of Mathematics at the Technical University Munich Deutschland. https://www.groups.ma.tum.de/statistics/personen/claudia-czado/forschung/lecture-slides/

[11] Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013) Regression Models. In: Fahrmeir, L., Kneib, T., Lang, S. and Marx, B., Eds., *Regression*, Springer, 21-72. https://doi.org/10.1007/978-3-642-34333-9_2

[12] Allen, M.P. (2004) Understanding Regression Analysis. Springer Science & Business Media.

[13] Brown, J.D. (2014) Linear Models in Matrix Form. Springer.

[14] Christensen, R. (1996) Analysis of Variance, Design, and Regression: Applied Statistical Methods. CRC Press.

[15] Christensen, R. (2018) Analysis of Variance, Design, and Regression: Linear Model-

ing for Unbalanced Data. Chapman and Hall/CRC.

[16] Horton, N.J. and Kleinman, K. (2015) Using R and RStudio for Data Management, Statistical Analysis, and Graphics. CRC Press.

[17] Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, **135**, 370-384. https://doi.org/10.2307/2344614

[18] Abraham, B. and Ledolter, J. (2006) Student Solutions Manual for Introduction to Regression Modeling. University of Iowa.

[19] Ricci, L. (2010) Adjusted-Squared Type Measure for Exponential Dispersion Models. *Statistics & Probability Letters*, **80**, 1365-1368. https://doi.org/10.1016/j.spl.2010.04.019

[20] McNeil, K.A., Newman, I. and Kelly, F.J. (1996) Testing Research Hypotheses with the General Linear Model. SIU Press.

[21] Seber, G.A. (2015) The Linear Model and Hypothesis. Springer.

[22] Vik, P. (2013) Regression, Anova, and the General Linear Model: A Statistics Primer. SAGE Publications.

[23] Lin, D.Y., Wei, L.J. and Ying, Z. (2002) Model-Checking Techniques Based on Cumulative Residuals. *Biometrics*, **58**, 1-12. https://doi.org/10.1111/j.0006-341x.2002.00001.x

[24] Osborne, J.W. and Waters, E. (2002) Four Assumptions of Multiple Regression that Researchers Should Always Test. *Practical Assessment, Research, and Evaluation*, **8**, Article 2.

[25] Lindsey, J.K. (2000) Applying Generalized Linear Models. Springer Science & Business Media.

[26] Farrar, D.E. and Glauber, R.R. (1967) Multicollinearity in Regression Analysis: The Problem Revisited. *The Review of Economics and Statistics*, **49**, 92-107. https://doi.org/10.2307/1937887

[27] Neter, J., Wasserman, W. and Kutner, M.H. (1983) Applied Linear Regression Models. Richard D. Irwin.

[28] Rusakov, O.V., Laskin, M.B. and Jaksumbaeva, O.I. (2015) Stochastic Pricing Model for the Real Estate Market: Formation of Log-Normal General Population. *Statistics and Economics*, No. 5, 116-127. https://doi.org/10.21686/2500-3925-2015-5-116-127

[29] Laskin, M.B. and Rusakov, O.V. (2023) Prediction of Distributions of Unit Prices for Real Estate Properties on the Basis of the Characteristics of Psi-Processes. *Business Informatics*, **17**, 7-24.

[30] Onumadu, U., Iyelobu, M., Akinde, M., Savadogo, S. and Yessoufou, B. (2025) Applications of Probability Distributions in Real Estate Market Analysis Using Rental Prices and Transaction Data in Major US Cities. *International Journal of Scientific Research in Information Technology*, **12**, 25-42.

[31] U.S. Department of Housing and Urban Development (2023) Worst Case Housing Needs: 2023 Report to Congress.

[32] Luca, S. (2023) Is Housing Price Distribution across Cities, Scale Invariant? Fractal Distribution of Settlements' House Prices as Signature of Self-Organized Complexity. *Chaos, Solitons & Fractals*, **174**, Article 113766. https://doi.org/10.1016/j.chaos.2023.113766